Univerza v Ljubljani
Fakulteta za računalništvo in informatiko

Darjan Oblak

Selekcija v skupinskih modelih z odločitvenimi drevesi

diplomsko delo na
univerzitetnem študiju
računalništva in informatike

Mentor: izr. prof. dr. Janez Demšar

Ljubljana, 2016

Fakulteta za računalništvo in informatiko izdaja naslednjo nalogo:

Skupinski modeli veljajo za najuspešnejše metode strojnega učenja, izvzemši novejše globoke nevronske mreže. Takšne modele sestavlja večje število notranjih modelov, med katerimi pa so tudi takšni, ki ne prispevajo k točnosti celotnega modela ali pa jo celo zmanjšujejo. Preučite obstoječe pristope k izboru notranjih modelov v skupinskih modelih. Razmislite o razlogih za njihovo uspešnost ali neuspešnost in na podlagi tega predlagajte morebitne nove, boljše postopke za ta namen.

Izjava o avtorstvu zaključnega dela

Spodaj podpisani Darjan Oblak, vpisna številka 63040115, avtor pisnega zaključnega dela študija z naslovom::

Selekcija v skupinskih modelih z odločitvenimi drevesi
(angl. *Decision Tree Ensemble Selection*)

IZJAVLJAM

1. da sem pisno zaključno delo študija izdelal samostojno pod mentorstvom izr. prof. dr. Janeza Demšarja;
2. da je tiskana oblika pisnega zaključnega dela študija istovetna elektronski obliki pisnega zaključnega dela študija;
3. da sem pridobil vsa potrebna dovoljenja za uporabo podatkov in avtorskih del v pisnem zaključnem delu študija in jih v pisnem zaključnem delu študija jasno označil;
4. da sem pri pripravi pisnega zaključnega dela študija ravnal v skladu z etičnimi načeli in, kjer je to potrebno, za raziskavo pridobil soglasje etične komisije;
5. soglašam, da se elektronska oblika pisnega zaključnega dela študija uporabi za preverjanje podobnosti vsebine z drugimi deli s programsko opremo za preverjanje podobnosti vsebine, ki je povezana s študijskim informacijskim sistemom članice;
6. da na UL neodplačno, neizključno, prostorsko in časovno neomejeno prenašam pravico shranitve avtorskega dela v elektronski obliki, pravico reproduciranja ter pravico dajanja pisnega zaključnega dela študija na voljo javnosti na svetovnem spletu preko Repozitorija UL;
7. dovoljujem objavo svojih osebnih podatkov, ki so navedeni v pisnem zaključnem delu študija in tej izjavi, skupaj z objavo pisnega zaključnega dela študija.

V Ljubljani, dne 24. avgusta 2016 Podpis avtorja:

Profesorju dr. Janezu Demšarju se iskreno zahvaljujem za nasvete, usmeritve ter odzivnost in prilagodljivost pri pisanju tega dela. Hvala tudi dr. Mihu Štajdoharju za svetovanje pri prvotno zastavljeni temi diplomskega dela.

Kazalo

Seznam uporabljenih kratic

kratica	angleško	slovensko
CART	Classification And Regression Tree	klasifikacijsko in regresijsko drevo
ECOC	Error Correcting Output Codes	izhodne kode za popravljanje napak
ET	Extremely Randomized Trees	ekstremno naključni gozdovi
ID3	Iterative Dichotomiser 3	-
MDSQ	Margin Distance Minimization	zmanjševanje robne razdalje
MeanD-M	Mean Decrease in Margin	zmanjšanje povprečja roba
MeanD-OM	Mean Decrease in OOB Margin	zmanjšanje povprečja OOB-roba
MinD-M	Decrease in Minimum Margin	zmanjšanje minimalnega roba
MinD-OM	Decrease in Minimum Margin	zmanjšanje minimalnega OOB-roba
MT	Margin Transformation	transformacija roba
OOB	Out-of-Bag	izven učnih podatkov
OVA	One-Versus-All	eden proti vsem
OVO	One-Versus-One	eden proti enemu
PRV	Parametrized Reference Vector	parametrizirani referenčni vektor
RF	Random Forest	naključni gozdovi
UCI	University of California, Irvine	Univerza v Kaliforniji, Irvine

Povzetek

Naslov: Selekcija v skupinskih modelih z odločitvenimi drevesi

Različne vrste skupinskih modelov se odlikujejo kot ene izmed uspešnejših metod strojnega učenja. Zaradi lepe lastnosti, ki jo imajo, da se točnost ob povečevanju števila notranjih modelov približuje asimptotični zgornji meji, imajo tudi slabost – velikost. V literaturi je moč zaslediti različne pristope, ki iščejo kompromis med velikostjo in točnostjo s postopkom selekcije. To pomeni, da v končni model uvrstijo le nekatere izmed generiranih notranjih modelov. Izkaže se, da na ta način ni možno le zmanjšati skupinskih modelov, temveč tudi povečati točnost. V tem delu metodam s selekcijo dodamo dva nova pristopa, ki za selekcijo uporabljata rob, ki ga modeli določajo na t. i. *out-of-bag* množici. Slednje je ključno pri majhnih podatkovnih množicah, saj to omogoča selekcijo brez izgube točnosti zaradi manjše učne množice. Metode ovrednotimo na 34 podatkovnih množicah za *bagging*, naključne gozdove in ekstremno naključne gozdove. Pri tem ugotovimo, da se v nekaterih primerih metode s selekcijo obnesejo statistično značilno bolje kot metode osnovnega skupinskega modela. V ostalih primerih metode s selekcijo uspešno zmanjšajo skupinski model in pri tem v povprečju ohranjajo točnost.

Ključne besede: skupinski modeli, odločitvena drevesa, selekcija v skupinskih modelih, rezanje skupinskih modelov, tanjšanje skupinskih modelov, bagging, naključni gozdovi, ekstremno naključni gozdovi.

Abstract

Title: Decision Tree Ensemble Selection

Ensemble models are well-known in machine learning for their accuracy. Their main quality, convergence towards an asymptotic upper limit as the number of internal models increases, is however partly counterbalanced by their large size. Existing studies show that posterior reduction of the number of models in the ensemble can be done without hurting – or with even increasing – the accuracy of the ensemble. The thesis introduces two new approaches to ensemble selection using the so-called "out-of-bag" set. Using such a selection set is important in case of small training sets where no data should be held out for learning in order to maintain high generalization accuracy of an ensemble. Both methods are evaluated on 34 datasets for bagging, random forest and extra decision trees. Some of the comparisons show that the selection model outperforms the base ensemble method in a statistically significant manner. The other confirm that the methods are able to reduce the size of ensembles while on average maintaining accuracy.

Keywords: ensemble models, decision trees, ensemble selection, ensemble pruning, ensemble thinning, bagging, random forest, extremely randomized trees.

Poglavje 1

Uvod

Klasifikacija obsega del problemov, s katerimi se ukvarja nadzorovano učenje, poddomena strojnega učenja. Naloga klasifikacije je na podlagi obstoječega znanja ter primerov z znanimi atributi in znanim razredom, ki mu pripadajo, določiti pripadajoči razred primerom, za katere poznamo le atribute, ne pa tudi razreda. To pomeni, da lahko z različnimi metodami določimo napovedni model, včasih imenovan tudi hipoteza. Model je diskretna funkcija, ki prostor atributov preslika v razred. Ključno je, da to počne čim bolj točno. Pogosto je prednost tudi razumljivost, saj ta omogoča boljši vpogled v domeno problema oz. relacije med atributi in razredi. Kadar nas zanima predvsem točnost modela, je pogosto potrebno poseči k bolj zapletenim metodam strojnega učenja.

Eden izmed takšnih bolj zapletenih pristopov so skupinski modeli, sestavljeni iz več posameznih modelov, ki skupaj glasujejo za ciljno napoved. Njihova glavna kvaliteta je predvsem doseganje visoke točnosti napovedi. Višja točnost je posledica različnosti hipotez, ki skupaj tvorijo zanesljivejšo napoved oz. s povprečjem napovedi dosežemo, da posamezne napake hipotez popravljajo druge hipoteze. Za doseganje visoke točnosti modela običajno potrebujemo več sto notranjih modelov.

V tem delu smo se omejili na klasifikacijske probleme, kot metode za reševanje le-teh pa smo obravnavali skupinske modele, ki kot notranje mo-

dele uporabljajo odločitvena drevesa. Ena izmed najbolj poznanih metod v tej skupini so naključni gozdovi, predstavili pa bomo tudi druge sorodne pristope. Naključnost v izgradnji nam omogoča, da so si notranji modeli med seboj različni. Zhou in dr. [47] so na primeru skupinskega modela nevronskih mrež kot eni prvih pokazali, da lahko s selekcijo notranjih modelov dosežemo manjši in hkrati boljši klasifikator. To je bila zadostna motivacija, da se je kasneje pojavilo še večje število drugih pristopov za selekcijo modelov na različnih vrstah skupinskih modelov.

Zanimalo nas je, kakšni pristopi so se pojavili v literaturi za selekcijo klasifikatorjev na skupinskih modelih z odločitvenimi drevesi. Na podlagi obstoječih teoretičnih in eksperimentalnih ugotovitev smo tudi sami predlagati in ovrednotili nov pristop k selekciji. Ker ima selekcija dva cilja, (čim) manjšo velikost in (čim) višjo točnost, se med sabo glede na prioritete razlikujejo tudi algoritmi. Sami smo si za lastne metode kot primarni cilj zadali skozi proces selekcije dosegati višjo klasifikacijsko točnost, kot jo ima neselektirani skupinski model. Na velikost dobljenega modela se nismo posebej osredotočali.

V poglavju 2 najprej predstavimo razloge za uspešnost skupinskih modelov in predstavimo nekaj pomembnih pojmov, povezanih z njimi. Sledi predstavitev vrst skupinskih modelov z odločitvenimi drevesi, opredelitev selekcije v skupinskih modelih in pregled obstoječih objav iz tega področja. V poglavju 3 predlagamo in na podlagi večjega števila množic ovrednotimo nove metode selekcije v skupinskih modelih. V poglavju 4 podamo sklepne ugotovitve in predloge za nadaljnje raziskovanje. Dodatek A podaja podrobnejše rezultate vrednotenja metod. V dodatku B prikazujemo parametre metod, kalibrirane na celotnih množicah podatkov.

Poglavje 2

Skupinski modeli

Marquis de Condorcet je leta 1785 v delu *Essay on the Application of Analysis to the Probability of Majority Decisions* predstavil t. i. Condorcetov teorem porote. V teoremu opisuje problem doseganja odločitve pri skupini glasujočih, kjer odločitev sprejmejo z večinskim glasom. Če predpostavimo, da so glasovi med sabo neodvisni in ima posamezni glasujoči verjetnost p za pravilno odločitev, potem velja:

- če je $p > 0.5$, se verjetnost za pravilno odločitev z dodajanjem glasujočih povečuje proti 1 in
- če je $p < 0.5$, se verjetnost za pravilno odločitev z dodajanjem glasujočih zmanjšuje proti 0.

Teorem ima nekatere omejitve, izpostaviti moramo predvsem pogoj neodvisnosti glasov, kar je težko, pogosto celo nemogoče, doseči. Druga omejitev je glasovanje med le dvema možnostima. V literaturi najdemo mnoge objave, ki se ukvarjajo z različnimi izpeljavami teorema, njegovo formulacijo razširjajo in obravnavajo tudi njegove omejitve. A teorem že v osnovi podaja intuitivno obrazložitev za večjo natančnost skupinskih modelov. Če sestavimo skupinski model iz raznolikih notranjih modelov, ki so med sabo dovolj neodvisni, s tem običajno raznoliki, in hkrati še dovolj točni, s tem povečamo verjetnost za pravilno napoved oz. dosežemo višjo klasifikacijsko točnost.

Dietterich [15] za večjo točnost skupinskih modelov v primerjavi s posameznimi modeli navaja tri glavne razloge, ki osmislijo skupinske modele z vidika omejitev pri iskanju optimalne hipoteze v (omejenem) prostoru hipotez:

- **Statistični razlog**: Včasih je učna množica nesorazmerno manjša od razpoložljivega prostora hipotez. Na takšni množici podatkov običajno več hipotez dosega podobno točnost. Z uporabo več modelov dobimo povprečje v prostoru hipotez in se s tem izognemo tveganju izbire posameznega modela, ki ima sicer na učni množici visoko točnost, a ima veliko napako glede na pravo hipotezo.
- **Računski razlog**: Mnogi algoritmi hipoteze oblikujejo na požrešni način in hipoteza lahko predstavlja nek lokalni optimum (globalnega pogosto ni mogoče doseči). Algoritem svoj izračun začne v neki začetni točki in izoblikuje hipotezo. Povprečje lokalno optimalnih hipotez je z večjo verjetnostjo bližje pravi hipotezi kot posamezna hipoteza.
- **Predstavitveni razlog**: Za končno množico učnih primerov obstaja tudi končna množica razpoložljivih hipotez. Zato pogosto ta množica sploh ne vsebuje prave hipoteze. Z (uteženim) glasovanjem večih hipotez lahko dosežemo razširitev prostora hipotez in se s tem približamo pravi hipotezi.

V nadaljevanju za skupinske modele najprej predstavimo pregled splošnih tehnik za gradnjo skupinskih modelov, nato pa podrobnejši pregled pristopov h gradnji skupinskih modelov, ki kot notranje modele uporabljajo odločitvena drevesa.

2.1 Pristranskost in varianca, raznolikost, rob

To, da so specifični skupinski modeli uspešni, je lahko pokazati, težja naloga pa je razložiti razloge za njihovo uspešnost in na podlagi njih morda predlagati možne izboljšave. Najdemo utemeljitve na podlagi koncepta roba, mer raznolikosti, razstavitve napake na pristranskost in varianco ter druge,

ponavadi povezane koncepte. Smiselno je torej, da teoretične ugotovitve na kratko povzamemo, saj na nekaterih temeljijo kasneje predstavljene metode selekcije.

Napako hipoteze lahko razstavimo na pristranskost (*bias*) in varianco (*variance*) [24]. Napaka hipoteze je na realnih problemih običajno vedno prisotna, saj težko zajamemo vse potrebne informacije problema, ki ga rešujemo. Pristranskost izvira iz učnega algoritma, varianca pa iz učnih podatkov. Za zmanjšanje pristranskosti moramo spremeniti učni algoritem. Pristranskost in varianca sta si nasprotujoči in minimiziranje ene pomeni naraščanje druge, zato iščemo optimalni kompromis med njima – takrat je napaka najmanjša.

Kadar je kompleksnost modela premajhna, takrat imamo nizko varianco in premajhno prilagajanje (*underfitting*) učnim podatkom, kadar pa prevelika, pa visoko varianco in s tem preveliko prilagajanje (*overfitting*) učni množici.

Pristranskosti na realnih problemih ne moremo izračunati, lahko jo le ocenimo na umetno zgeneriranih problemih [24]. Običajno jo srečamo v kontekstu regresijskih problemov. Takšno razstavljanje napake na pristranskost in varianco prikaže Louppe v [28]. Na sliki 2.1 pa ponazorimo napako učne in testne množice v odvisnosti od kompleksnosti modela (globine odločitvenega drevesa). Poudarjeni krivulji predstavljata povprečje napake 50 različnih odločitvenih dreves pri 50 naključnih delitvah na učno in testno množico. Navpična črta označuje optimalni kompromis med pristranskostjo in varianco.

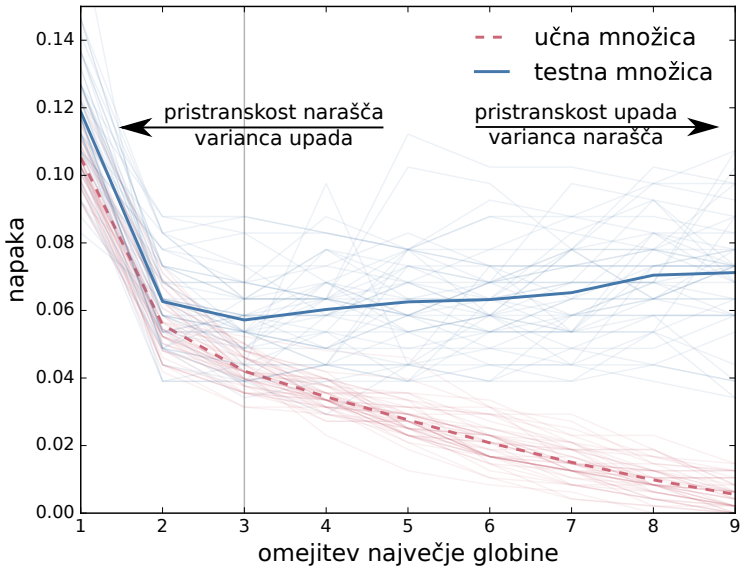
Schapire in dr. [41] definirajo rob kot razliko med številom glasov za pravilni razred in številom glasov tistega izmed ostalih razredov, ki je prejel največ glasov (s predpostavko, da uporabljamo glasovanje). Rob običajno obravnavamo normaliziranega med vrednosti -1 in 1. V omenjeneni objavi analizirajo napovedno napako na učni in testni množici ter distribucijo robov na učni množici za *boosting* in ga med drugim primerjajo z *baggingom*. S pomočjo distribucije robov razložijo razloge za uspešnost *boostinga*. Pokažejo namreč, da je možno na podlagi distribucije robov oceniti zgornjo mejo napo-

vedne napake modela. Poudarijo, da je sicer ta ocena preveč pesimistična za natančnejše ocenjevanje napake, a vendarle iz tega izhaja pomen vloge roba v povezavi s točnostjo. Svoje ugotovitve povežejo tudi s pristranskostjo in varianco. Na tem mestu naj omenimo, da bomo v nekoliko drugačnem kontekstu enako definicijo roba uporabljali tudi v naših metodah v povezavi z *out-of-bag* (OOB) množico [7], kjer je odstotek pozitivnih robov pravzaprav ocena točnosti modela. Seveda slednje velja le pred selekcijo, po selekciji moramo OOB-množico, v kolikor smo jo uporabili za selekcijo, smatrati tudi kot učno množico in za ocenjevanje točnosti potrebujemo novo neodvisno množico.

Kuncheva in Whitaker v [25] sta na podlagi različnih mer raznolikosti predstavila podrobnejšo analizo povezave med raznolikostjo in točnostjo skupinskih modelov. Najdeta le šibko povezavo med raznolikostjo in točnostjo modelov v splošnem, čeprav je močna korelacija prisotna v nekaterih posebnih primerih. Zaključita, da je zato uporabnost mer raznolikosti kot indikatorja za točnost modelov vprašljiva. Pri gradnji modelov je raznolikost sicer potreben pogoj za komplementarnost modelov, ni pa sama po sebi garancija za točnost.

Za raznolikost Tang in dr. [42] pojasnijo, da kljub temu, da gre za zelo pogost pojem v skupinskih modelih, ne obstaja enotna definicija, kaj raznolikost sploh je, saj gre za zahteven problem. Zato srečamo večje število različnih mer raznolikosti. V objavi se pod pojmom raznolikost nanašajo na 6 v drugih objavah predstavljenih mer raznolikosti. V skupinskih modelih pravimo, da model lahko implicitno stremi k raznolikosti, takšni so recimo skupinski modeli, predstavljeni poglavju 2.3. Lahko pa raznolikost iščejo eksplicitno skozi mere raznolikosti. Pokazali so, da si z eksplicitnim iskanjem raznolikosti ne moremo obetati konsistentno dobrih rezultatov. Na iskanje raznolikosti namreč lahko gledamo kot na optimiziranje roba v smislu vloge pri zgornji meji napovedne napake, ta pa se ne povečuje monotono glede na raznolikost.

Ko in Sabourin [23] razširita koncept raznolikosti tako, da združita po-



Slika 2.1: Napaka na učni in testni množici v odvisnosti od globine odločitvenega drevesa za podatkovno množico bcw (idejno povzeto iz [20] in [28]).

znane mere raznolikosti v funkcijo, ki hkrati upošteva natančnost individualnih modelov in raznolikost. V nasprotju z dotedanjimi raziskavami, ki so pokazale le šibko korelacijo med raznolikostjo in točnostjo skupinskega modela, uspeta pokazati močno korelacijo med točnostjo skupinskega modela in razširjenim konceptom raznolikosti, združene s točnostjo posameznih klasifikatorjev. Pri tem imajo novo predlagane mere raznolikosti tudi večjo medsebojno korelacijo. Zaključita, da je smiselno pri gradnji skupinskih modelov hkrati upoštevati tako raznolikost kot tudi točnost posameznih modelov. Nepovezano tudi Brown in Kuncheva [10] pojmu raznolikosti dodata komponento individualne natančnosti modelov. Napako večinskega glasovanja predstavita kot rezultat treh komponent, "dobre" in "slabe" raznolikosti ter točnosti posameznega modela. Dobra raznolikost zmanjšuje napovedno napako, slaba pa jo povečuje.

2.2 Tehnike za gradnjo skupinskih modelov

Predlagane modifikacije skupinskih modelov, ki bodo predstavljene v naslednjem poglavju, se nanašajo na manjšo podmnožico skupinskih modelov. Za lažje razumevanje teoretičnega ozadja pa je smiselno, da to podmnožico najprej umestimo v kontekst. Zaradi izjemne razvejanosti področja pregled zagotovo ni celovit, se pa skuša osredotočiti na bolj poznane skupinske modele, ki jih srečamo v strojnem učenju.

Kot že omenjeno je ena od ključnih lastnosti skupinskih modelov raznolikost. Za zagotavljanje te potrebujemo ustrezne mehanizme. Bagheri v [2] predstavi štiri ključne principe pri gradnji skupinskih modelov:

- **Manipulacija množice primerov (*subsample approach*)**: Za učenje posameznega modela se uporabi transformirana množica primerov. Primeri so lahko izbrani s ponavljanjem ali brez njega. Lahko je zgrajena naključno ali pa načrtno z namenom osredotočanja na določen del celotne množice. Lahko gre za podmnožico osnovnih primerov, lahko pa vsebuje umetno zgrajene primere na podlagi prvotne množice.
- **Manipulacija atributov (*subspace approach*)**: Za izgradnjo modela v celoti ali pa samo v posameznih fazah izgradnje modela se uporabi podmnožico ali transformirano množico atributov. Atribute je možno razvrstiti v skupine glede na naravo problema, lahko jih izberemo naključno ali z določeno hevristiko.
- **Manipulacija razredov (*subclass approach*)**: Večina pristopov se osredotoča na t. i. binarizacijo razredov, kjer namesto enega problema klasifikacije v več razredov rešujemo več problemov binarne klasifikacije med dvema posameznima podmnožicama razredov.
- **Manipulacija modelov (*learner manipulation approach*)**: Uporabi se različne vrste modelov in/ali se posamezni vrsti uporabljenega modela spreminja parametre.

V naslednjem podpoglavju bomo v zgornje kategorije razvrstili bolj poznane skupinske modele z odločitvenimi drevesi. Prej pa povzamemo še širšo

opredelitev pristopov h gradnji skupinskih modelov, ki jo poda Rokach [40]:

- **Odvisni pristopi (*dependent frameworks*)**: Posamezni notranji model se zgradi na podlagi prejšnjih izgrajenih modelov. V okviru teh poznamo:
  - Inkrementalno učenje (*incremental batch learning*): Prejšnji izgrajeni model se uporabi kot predznanje pri gradnji naslednjega modela. Zadnji izgrajeni model je izbran kot končni model.
  - Izbiranje učne množice na podlagi modelov (*model-guided instance selection*): Pri gradnji naslednjega modela, ki se doda množici zgrajenih, se uporabijo vsi prejšnji modeli, tako da se na podlagi teh izbere učno množico za gradnjo naslednjega modela. Običajno to pomeni, da se naslednji modeli učijo samo na napačno klasificiranih primerih. Najbolj znan pristop v tem sklopu je *boosting*.
- **Neodvisni pristopi (*independent frameworks*)**: Notranji modeli se zgradijo neodvisno eden od drugega, kar omogoča tudi paralelizacijo. Lahko so iste ali različnih vrst. Tipični predstavnik je *bagging*.

V tem delu se bomo osredotočili na neodvisne pristope. Vsi notranji modeli se zgradijo vnaprej, nato pa se na teh izvede postopek selekcije in dobimo manjši skupinski model s podmnožico modelov.

Rokach [40] definira tudi ključne komponente pri gradnji skupinskega modela (delni povzetek):

- **Učna množica (*training set*)**: Primeri z naborom atributov A in ciljno spremenljivko razreda y.
- **Algoritem za izgradnjo modela (*base inducer*)**: Algoritem za izgradnjo izbrane vrste modelov. Na podlagi učne množice S algoritem I zgradi model M.
- **Generator raznolikosti (*diversity generator*)**: Komponenta, ki zagotavlja, da so generirani modeli raznoliki. Principe za zagotavljanje raznolikosti smo že povzeli po Bagheriju [2].

- **Kombinator (*combiner*)**: Napovedi notranjih modelov kombinira v napoved skupinskega modela. Poznamo:
  - **Uteževanje**:
    - **Večinsko glasovanje**: Izbran je tisti razred, za katerega glasuje največ klasifikatorjev.
    - **Po točnosti**: Uteževanje glasov glede na točnost na validacijski množici.
    - **Povprečje verjetnosti**: Izbran je razred z najvišjim povprečjem pogojnih verjetnosti posameznih klasifikatorjev.
    - ***Vogging* (*variance optimized bagging*)**: Izbira takšne linearne kombinacije klasifikatorjev, da se kar najbolj zmanjša varianco in se pri tem ohrani točnost.
    - **Glede na entropijo**: Vsak klasifikator dobi utež, obratno sorazmerno z entropijo vektorja klasifikacij.
    - **Gostota učnega prostora**: Osnovna predpostavka je, da so bili klasifikatorji naučeni na različnih podmnožicah podatkov. Glas posameznega klasifikatorja za primer x se uteži glede na verjetnost, da je bil klasifikator naučen na podatkih, ki ustrezajo prostoru, kamor spada x.
  - **Skladanje (*stacking*, metakombiniranje)**: Običajno se uporablja za kombiniranje različnih vrst modelov. Osnovna ideja je zgraditi meta-učno množico, ki ima namesto atributov osnovne učne množice predvidene klasifikacije modelov, razred pa ostane isti. Zgradi se metaklasifikator, ki kombinira napovedi osnovnih modelov v končno napoved. Pokazano je bilo, da ima takšen model lahko višjo točnost kot izbira najboljšega izmed klasifikatorjev v prečnem preverjanju [16].

Kot posebno vrsto uteževanja lahko razumemo tudi v poglavju 2.4 predstavljeno selekcijo, na katero se osredotočamo v tem delu. Selekcijo bi namreč lahko definirali tudi tako, da nekaterim od dreves priredimo utež 0, drugim pa 1, če gre za selekcijo brez ponavljanja, ali tudi druga cela števila ≥ 1, če

Tabela 2.1: Skupinski modeli glede na uporabljene principe gradnje

	manipulacija množice primerov	manipulacija atributov	manipulacija razredov	manipulacija modelov
bagging, *boosting*, DECORATE, *pasting*	•			•
naključni gozdovi, rotacijski gozdovi, *random patches*	•	•		•
naključni podprostor, ekstremno naključni gozdovi		•		•
eden-proti-enemu (OVO), eden-proti-vsem (OVA), kode za popravljanje izhodnih napak (ECOC)			•	•

gre za selekcijo s ponavljanjem. V tuji literaturi zasledimo omenjeno tehniko pod izrazi *pruning*, *selection* in *thinning*.

2.3 Skupinski modeli z odločitvenimi drevesi

Bagheri [2] predstavljene principe poveže z znanimi vrstami skupinskih modelov, kar deloma povzemamo (in z za naše delo relevantnimi metodami dopolnjujemo) v tabeli 2.1.

2.3.1 Odločitvena drevesa

Odločitvena drevesa so sestavljena iz notranjih vozlišč, vej in listov. Vozlišča predstavljajo atribute, veje podmnožice vrednosti atributov, listi pa ciljne razrede. Pot od korena do lista predstavlja odločitveno pravilo. Preprosto metodo za gradnjo odločitvenih dreves sta že 1963 predstavila Morgan in Sonquist [33]. Skozi čas so se pojavljali izboljšani pristopi. Med danes najbolj poznane sodijo ID3 [36], C4.5 [38] in CART (Classification and Regression Trees) [5]. Izraz CART pogosto srečamo tudi kot splošni akronim za odločitvena drevesa.

V eksperimentalnem delu uporabljamo programsko orodje scikit-learn[1] [34] in s tem optimizirano verzijo CART algoritma za drevesa v skupinskih modelih.

2.3.2 Pasting

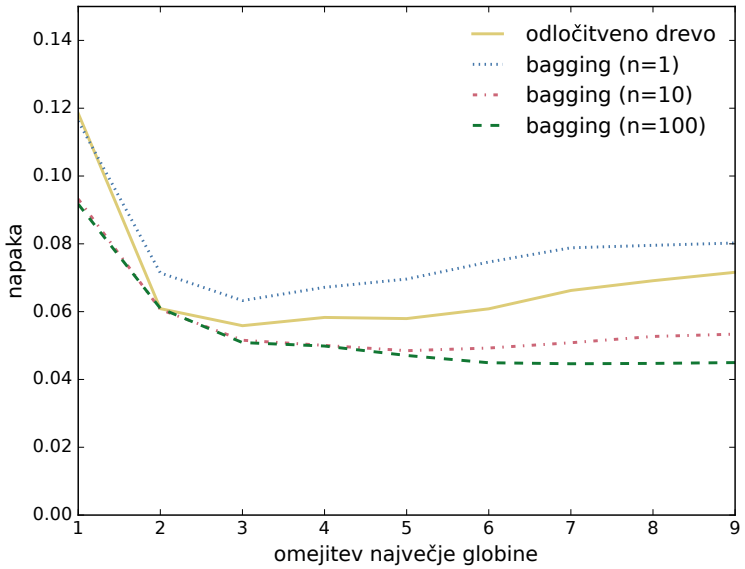
Breiman [8] metodo *pasting* predlaga primarno za reševanje problemov, kjer imamo velike količine podatkov, ki jih težko vse hkrati shranjujemo v hitrem pomnilniku. Iz celotne množice podatkov se za učenje vsakič uporabi le podmnožica vseh podatkov. Podmnožica podatkov se lahko izbere naključno ali pa glede na pomembnost s prioriteto na primerih, za katere ocenjujemo, da imajo večjo verjetnost napačne klasifikacije. Modele gradimo, dokler se klasifikacijska napaka zmanjšuje – za to lahko uporabimo testno množico podatkov ali pa OOB-množico.

2.3.3 Bagging

Že pred *pastingom* je Breiman [6] razvil metodo večkratnega učenja, ki naključne podmnožice izbira s ponavljanjem primerov, kar imenuje *bagging*. Če celotna množica obsega N primerov, s ponavljanjem izvlečene učne množice za gradnjo posameznih modelov pa ravno tako N primerov, potem posamezni primer v vsaki od množic nastopa z verjetnostjo ≈ 0.632. Ker v posamezni učni množici za gradnjo modela ne nastopajo vsi primeri in ponavljanje doda nekaj dodatne nestabilnosti, dobimo tudi bolj raznolike (angl. *diverse*) modele. Ti so posamezno manj točni, a kot skupinski model pogosto dosegajo večjo točnost. *Bagging* poveča varianco notranjih modelov, a običajno s povprečenjem na skupinskem modelu bolj zniža skupno napako, ki izvira iz variance, kot poviša prispevek napake iz pristranskosti. Quinlan [37] je pokazal, da *boosting* dosega v povprečju boljšo točnost kot *bagging*, a je bolj občutljiv na šum.

V poglavju 2.1, kjer obravnavamo pristranskost in varianco, smo na sliki 2.1

[1]Dostopno na: http://scikit-learn.org/stable/, verzija 0.17.1



Slika 2.2: Točnosti napovedi metode *bagging* v primerjavi z odločitvenim drevesom v odvisnosti od globine dreves na podatkovni množici bcw.

prikazali, kako na napovedno napako vpliva naraščajoča kompleksnost modela oz. naraščajoča varianca. Tej napaki smo na sliki 2.2 dodali še napako *bagginga* ob različnem številu notranjih dreves. Vidimo lahko, da ima posamezno drevo pri *baggingu* v povprečju večjo napako kot odločitveno drevo, zgrajeno na celotni množici podatkov. Pri povečevanju števila notranjih modelov pa *bagging* zmanjša varianco skupinskega modela in tako doseže višjo točnost.

2.3.4 Naključna podmnožica atributov v vsakem vozlišču gradnje drevesa

Amit in dr. [1] pri reševanju problema prepoznave pisave uporabijo pristop h grajenju odločitvenih dreves, kjer pri vsakem iskanju najboljšega atributa za binarno delitev v vozlišču uporabimo le naključno podmnožico razpoložljivih atributov. Velikost te podmnožice se v literaturi običajno označuje s K, $max\_features$ ali $mtry$. Izkaže se, da tako nastanejo drevesa, ki so si

po strukturi različna, a hkrati še vedno dokaj točna. Višjo točnost lahko dosežemo s povprečenjem večih dreves.

2.3.5 Naključni podprostor

Ho [22] je na podlagi prejšnjega pristopa izpeljala metodo naključnega podprostora (*random subspace*). Razlika je v tem, da se naključna podmnožica atributov izbere samo enkrat, pred gradnjo vsakega posameznega drevesa, ne na vsakem vozlišču posebej. Na ta način uspemo zgraditi bolj raznolike modele, ki kot skupinski model lahko dosegajo višjo točnost. Hkrati lahko služi na podoben način kot pasting – zmanjšuje zahteve po razpoložljivem hitrem pomnilniku.

2.3.6 Naključni gozdovi (random forest, RF)

Breiman [9] je v ideji naključnih gozdov združil metodo *bagging* in izbiro naključne podmnožice atributov v vsakem vozlišču gradnje drevesa. V vlogi generatorja raznolikosti se tako hkrati uporabita dva mehanizma. Posameznih dreves pred ali po gradnji ne režemo. Z *baggingom* pridobimo točnost, hkrati pa tudi OOB-oceno napake. Metoda dosega primerljivo točnost z *boostingom*, pri čemer je manj občutljiva na šum. Ključna parametra pri uporabi metode sta število zgrajenih modelov $n\_estimators$ in velikost naključnih podmnožic atributov K. S slednjim lahko, ravno tako kot pri prejšnjih dveh metodah, nadzorujemo razmerje med varianco in pristranskostjo [28]. Scikit-learn v nasprotju z izvornim predlaganim algoritmom uporablja povprečje verjetnosti notranjih klasifikatorjev in ne večinskega glasovanja.

2.3.7 Ekstremno naključni gozdovi (*Extremely Randomized Trees*, ET)

Geurts in dr. [18] predlagajo metodo, podobno RF, ki ne uporablja vzorčenja s ponavljanjem oz. *bagginga* za izbiro učne množice za posamezni model,

temveč se vsak model zgradi na celotni učni množici. To komponento generatorja raznolikosti v metodi nadomesti naključna izbira delitvenih točk za binarno delitev pri atributih. ET uspe, podobno kot RF, v primerjavi z odločitvenim drevesom v večini primerov pri majhnem povečanju pristranskosti s povprečenjem modelov znatno bolj znižati varianco. Eksperimentalno se izkaže, da ET v povprečju dosega boljše rezultate kot RF [18, 28].

2.4 Selekcija v skupinskih modelih

V nadaljevanju opredelimo vrste selekcije in povezave nekaterih značilnosti skupinskih modelov v povezavi s točnostjo napovedi. Sledi delni pregled obstoječih metod selekcije.

Zhou [46] v obsežnem delu, ki celovito predstavlja področje skupinskih modelov, definira tri glavne kategorije selekcije modelov:

- **Selekcija na osnovi razvrščanja**: Modele se razvrsti na podlagi izbranega kriterija. V končni izbor pa je izbran le del vseh modelov glede na razvrstitev.
- **Selekcija na osnovi gručenja**: Množico modelov se razdeli v gruče, pri čemer v posamezno gručo spadajo modeli s podobnimi karakteristikami, gruče pa so med sabo raznolike. Prototip posamezne gruče se uvrsti v množico izbranih modelov.
- **Selekcija na osnovi optimizacije**: Selekcijo dreves definira kot optimizacijski problem, katerega cilj je maksimizirati ali minimizirati izbrano karakteristiko skupine in tako najti podmnožico modelov, ki skupaj kot celota dosegajo visoko klasifikacijsko točnost.

Meje med kategorijami niso ostre in nekatere metode je težko razvrstiti v eno samo kategorijo, saj uporabljajo kombinirane pristope. Selekcijo opiše tudi kot posebno vrsto skladanja. Če namreč skladanje temelji na metamodelu, ki kombinira posamezne modele v skupno napoved, lahko na selekcijo gledamo tudi iz tega vidika – metamodel je v tem primeru selekcijski postopek.

Potrebno je razlikovati med selekcijo, ki smo jo ravnokar definirali, in pa tehniko, ki jo izvajajo odvisni pristopi gradnje skupinskih modelov. Slednji v posameznih fazah lahko zavržejo posamezne zgrajene modele, če ti ne ustrezajo določenim kriterijem (npr. pri AdaBoost, če je točnost posameznega klasifikatorja < 0.5). Kljub temu, da se pojavljajo mešani pristopi, ki selekcijo uporabljajo že tekom posameznih faz odvisnih skupinskih metod, pa po osnovni definiciji do selekcije pride šele v končni fazi, ko so že generirani vsi modeli in se po sami selekciji ne generira dodatnih modelov [46].

V predstavitvi metod selekcije iz literature dajemo poudarek na bolj poznane in takšne, ki so sorodne predlaganim v tem diplomskem delu. Selekcija na osnovi gručenja je manj pogost pristop, hkrati pa tudi najmanj soroden predlaganim v tem delu, zato ga ne predstavljamo posebej. Izognemo se tudi poglobitvi v kompleksne tehnike selekcije na osnovi optimizacije, ki eksplicitno iščejo podmnožico, ki bi imela čim boljšo napovedno točnost. Krajši pregled je moč najti v [31, 35]. Qian in dr. [35] predstavijo tudi tudi teoretične relacije med pristopi na osnovi optimizacije in pristopi na osnovi razvrščanja, predvsem z vidika optimalnosti rešitve in časovne zahtevnosti.

Metode selekcije na osnovi razvrščanja se poslužujejo izbire funkcije, ki jo v procesu razvrščanja maksimizirajo ali minimizirajo. Izbira modelov v selektirani skupinski model nato temelji na izbiri določenega odstotka modelov glede na razvrstitev. Učinkovitost takšnih metod za klasifikacijo najdemo empirično potrjeno v [31], za regresijo, ki jo v tem delu sicer ne obravnavamo, pa v [21].

Glavna težava, na katero smo naleteli pri pregledovanju literature, se zdi velika specifičnost eksperimentalnega dela objav. Neredko je uporabljen majhen vzorec podatkovnih množic in/ali na samo ena vrsta skupinskega modela. Nekatere metode ovrednoti več različnih objav, a pri tem rezultati niso vedno konsistentni. Podrobnejšo analizo in razloge za to bomo podali v sledečem poglavju, v tem poglavju pa v nadaljevanju sledi predstavitev principov delovanja nekaterih metod na osnovi razvrščanja.

2.4.1 Selekcija z zmanjševanjem napake (*reduce-error pruning*, RE)

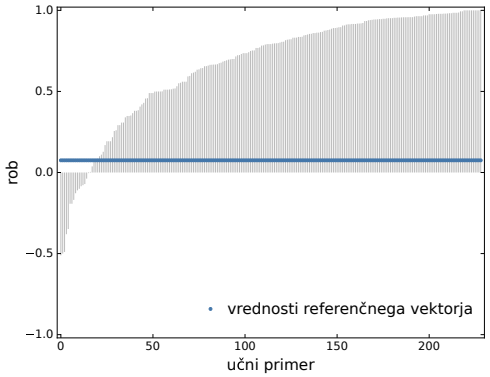
Margineantu in Dietterich [29] predstavita metodo, temelječo na zmanjševanju napake napovedi z uporabo vzvratnega prilagajanja *backfitting* [17] na primeru *boostinga*. Na podlagi te Martínez-Muñoz in Suárez [32] predlagata poenostavljeno različico brez vzvratnega prilagajanja za neodvisne pristope gradnje skupinskih modelov in v [30] tudi navedeta, da je računsko bistveno manj zahtevna, pri čemer se pri *baggingu* obnese enako dobro, kot če bi uporabili vzvratno prilagajanje.

Metodo lahko uvrstimo med pristope na osnovi razvrščanja. V algoritmu brez vzvratnega prilagajanja je v vsakem koraku algoritma v množico izbranih dodan tisti model, ki zagotavlja čim manjšo napako napovedi na selekcijski množici.

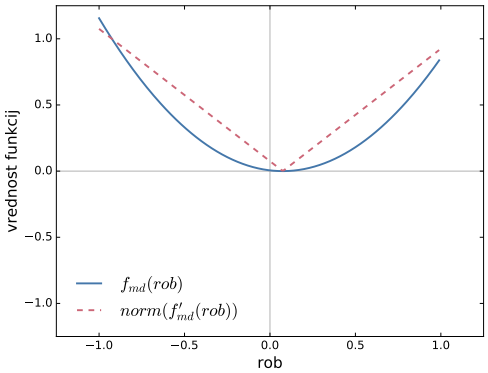
2.4.2 Selekcija na osnovi distribucije roba

Rob, na katerega se nanaša metoda, smo že opisali v poglavju 2.1. Vrednosti roba za posamezne primere spadajo med -1 in 1. Martínez-Muñoz in Suárez [32] predstavita pristop k selekciji skupinskih modelov, ki temelji na minimizaciji robne razdalje do referenčnega vektorja. Metodo poimenujeta selekcija z zmanjševanjem robne razdalje (*margin distance minimization*, MDSQ). Imamo selekcijsko množico N_{sel} in za vsak posamezni klasifikator definiran $|Nsel|$-dimenzionalni vektor c_t, katerega i-ta vrednost znaša 1, če izbrana množica pravilno klasificira primer i, in -1, če ga ne. Povprečje istoležnih komponent po vseh vektorjih nam poda nov vektor $\langle c \rangle$, za katerega želimo, da leži čimbližje referenčnemu vektorju o, katerega vse komponente so enake vrednosti p. Izbrana vrednost p je majhna, tipično $p \in (0.05, 0.25)$. Algoritem korakoma izbira modele tako, da vsakič v množico N_{sel} doda tisti model, ki kar najbolj minimizira kvadratno razdaljo med $\langle c \rangle$ in o.

Delovanje metode lahko orišemo s sliko 2.3, kjer je prikazan referenčni vektor in rob za primere učne množice urejene po velikosti roba. Posamezni



Slika 2.3: Lega referenčnega vektorja pri metodi MDSQ in primer distribucije roba za učne primere podatkovne množice *breast-cancer*.



Slika 2.4: Drugačna predstavitev pristopa minizacije funkcije robne razdalje pri metodi MDSQ ($p = 0.075$).

modeli so nato v algoritmu požrešno dodajani v množico, ki postopoma kar najbolj minimizira funkcijo robne razdalje do referenčnega vektorja. Opozoriti moramo, da je prikazana distribucija robov značilna za neporezana drevesa, medtem ko v [31] ugotovijo, da metoda dobro deluje le pri porezanih drevesih. K temu vprašanju se bomo vrnili v sledečem poglavju. Ker ima MDSQ to lastnost, da so vse vrednosti referenčnega vektorja enake, ga lahko prevedemo v obliko predstavitve na sliki 2.4. Modra krivulja predstavlja funkcijo robne razdalje, ki jo minimiziramo. Prekinjena krivulja označuje med vrednosti 0 in 1 normaliziran odvod funkcije in s tem utež za "popravljanje" roba posameznega primera pri iskanju naslednjega najboljšega modela, ki ga želimo vključiti v skupinski model. To pomeni, da se z oddaljenostjo od referenčne točke utež za popravljanje posameznega primera linearno povečuje.

Martinez in dr. kasneje v [31] predlagajo izboljšano različico metode, ki uporablja premični referenčni vektor o, definiran z vrednostjo $p = 2\sqrt{2u}/T$, kjer je u velikost podmnožice modelov, ki bo izbrana po trenutnem koraku, T pa število vseh generiranih modelov. Metodo ovrednotijo na *baggingu* z uporabo porezanih dreves, pri čemer selekcija poteka na učni množici.

Yang in dr. [45] so selekcijo, temelječo na osnovi distribucije roba, analizirali na primeru naključnih gozdov. Modelov ne izbirajo v množico izbranih, temveč izločajo iz celotne množice. Primerjajo tako selekcijo na celotni učni množici kot tudi samo na OOB-množici. V primeru slednjega se za posamezni primer upoštevajo le glasovi tistih modelov, za katere primer ni bil uvrščen v učno množico. Predpostavijo, da je optimalna distribucija roba takšna, da imamo bodisi največjo povprečno vrednost robov, ali pa, da imamo čim višjo vrednost minimalnega roba. V vsakem koraku selekcije torej izločijo model z najmanjšo vrednostjo ene od funkcij:

- povprečno zmanjšanje roba na celotni učni množici (MeanD-M),
- zmanjšanje minimalnega roba na celotni učni množici (MinD-M),
- povprečno zmanjšanje roba na OOB-učni množici (MeanD-OM) in
- zmanjšanje minimalnega roba na OOB-učni množici (MinD-OM).

Kot najboljša metoda se izkaže MeanD-M.

Še drugačne variante selekcije z optimizacijo roba najdemo v [44] in [19].

2.4.3 Selekcija na osnovi mer raznolikosti

Predstavili smo že ugotovitve iz literature, kjer so analizirali (ne)učinkovitost uporabe mer za povečevanje točnosti napovedi. Pojavili so se pristopi, ki so to vseeno poizkušali, a so se empirično izkazali za relativno neučinkovite. Selekcijo z nekaterimi merami sicer analizirajo Banfield in dr. [3], a uspešnost pristopov prikažejo le v primerjavi z naključno selekcijo. Šele kasnejše objave so prišle do učinkovitih pristopov, kjer poleg raznolikosti upoštevajo še točnost posameznih modelov. Tako na podlagi ugotovitev iz [23] in [10] selekcijo v skupinskih modelih obravnavajo Bhatnagar in dr. [4].

Poglavje 3

Selekcija z uporabo roba

3.1 Predlagane metode

Obe predlagani metodi v tem delu spadata v kategorijo selekcije na osnovi razvrščanja in temeljita na optimizaciji distribucije roba. Bistvena razlika z obstoječimi metodami pa je v tem, da se optimizacija nanaša na vrednosti t. i. OOB-roba. OOB-rob je rob, ki ga izračunamo tako, da za posamezni primer uporabimo le glasove modelov, pri katerih primer ni nastopal v učni množici. To je torej možno pri metodah razmnoževanja učnih primerov. ET sicer ni takšna metoda, a v naslednjem poglavju predstavimo, kako smo to oviro zaobšli.

OOB-rob so za selekcijo, kot že omenjeno, uporabili v [45], a manj uspešno kot pri uporabi roba učne množice kot celote. Kljub temu smo ocenili, da bi bilo možno OOB-rob koristno uporabiti, če upoštevamo nekatere omejitve, ki jih ima. Prva pomembna ugotovitev izhaja iz dejstva, da je za ocenjevanje točnosti posameznega primera na voljo le približno 37 % vseh modelov [7]. To dejstvo igra pri selekciji na podlagi OOB-roba pomembno vlogo. Če namreč selekcijo izvajamo na skupinskih modelih z že sicer majhnim številom modelov, to pomeni, da je število glasov za ocenjevanje OOB-roba za posamezni primer relativno majhno in iz tega možna izhajajoča napaka večja.

Druga ugotovitev, da je OOB-ocena je nekoliko pesimistična [12, 11], je

verjetno manj pomembna, a jo je vseeno smiselno omeniti in se je zavedati pri razvoju metod. V slednjih objavah sicer najdemo predloge za korekcijo OOB ocene, s čimer bi verjetno lahko prišli tudi do korekcije vrednosti robov, na podlagi katerih izvajamo selekcijo. A pristranskosti roba nismo posvečali posebne pozornosti, saj smo ocenili, da z izločanjem modelov na podlagi samega OOB-roba v vsakem primeru izgubimo točnost OOB-ocene kot ocene točnosti skupinskega modela, zato se zdijo predhodne korekcije roba nesmiselno delo.

Za manjšo uspešnost metod, ki uporabljajo OOB-rob v [45], vidimo dva možna vzroka. Prvi je majhno število notranjih modelov (100), iz česar sledi večja nezanesljivost robnih vrednosti. Drugi pa je neustreznost samih metod za uporabo na OOB-robu. Potencialna težava pri maksimizaciji povprečnega roba je odsotnost mehanizma, ki bi preprečil izločitev takih modelov, ki komajda pozitiven rob posameznih primerov spremenijo v negativnega in s tem povečajo možnost za izgubo natančnosti. Pri drugi metodi, maksimizaciji minimalnega roba, pa ima lahko znaten vpliv šum, prisoten v podatkih.

Definicija iz [46], da je selekcija posebne vrste skladanje, se zdi še najboljši opis sledečih metod. Da bodo sploh uporabne, predvidevamo zaradi dejstva, da imamo pri metodah z OOB-množico v obliki posameznega učnega primera za približno 37 % modelov na voljo še nevideno informacijo. Ti primeri so bili z razlogi, pojasnjenimi v predhodnem poglavju, izključeni iz učne množice za posamezne modele. Hkrati intuitivno menimo, da nam OOB-rob te primere uvrsti v tri skupine, med katerimi meje niso jasno začrtane. Prva skupina so primeri z visokim robom, druga primeri z robom blizu vrednosti 0, tretja pa močneje negativni primeri. Sklepamo, da bi si za prvo skupino, primere z visokim robom, lahko brez škode za natančnost privoščili izgubiti nekaj pravilnih glasov, če bi na ta račun uspeli izločiti tudi nekaj glasov za primere z robom blizu vrednosti 0, ki morda niso šum in so lahko napačno klasificirani. Učne primere, iz katerih lahko pridobimo OOB-oceno, na ta način "žrtvujemo" za dodatno metaučenje.

Pri predlaganih metodah modele iz skupinskega modela izločamo tako kot

v [45], torej ravno obratno kot v [32, 31, 35], kjer se modele dodaja v množico izbranih. Razlog za izbiro takega pristopa je ta, da selekcije ne izvajamo na celotni učni množici ali posebni selekcijski množici, temveč vsak klasifikator prispeva napovedno vrednost le k primerom iz svoje OOB-množice.

Za zmanjševanje časovne zahtevnosti algoritma uvedemo dodatno prilagoditev – najboljšega kandidata za izločitev ne izberemo vedno optimalno med vsemi razpoložljivimi modeli, temveč v zanki izločimo vsak model, za katerega se funkcija robne razdalje poveča ali ostane enaka. Z izločanjem prenehamo, če ni možno izločiti nobenega modela več.

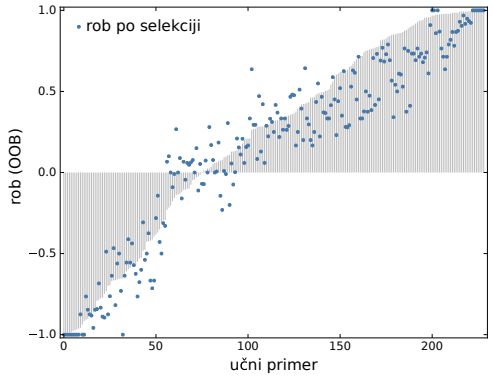
3.1.1 Selekcija s transformacija roba (MT)

V poglavju 2.4.2 smo opisali metodo MDSQ, ki minimizira funkcijo robne razdalje. Pristop smo preizkusili na OOB-množici, a se je izkazalo, da uspešnost močno niha glede na vrednost o. Na sliki 2.3 smo prikazali rob celotne učne množice. Na sliki 3.1 za primerjavo prikazujemo OOB-rob za isto množico. Z modro pa je označen rob po selekciji, pri čemer je razvidno, da prej pozitiven OOB-rob nekaterih primerov postane negativen. To ocenjujemo kot problematično in vir slabše točnosti selektiranega skupinskega modela.

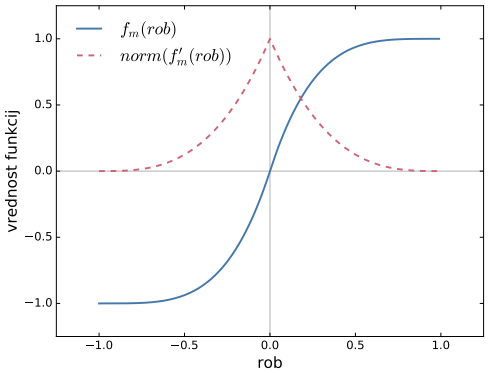
Težavo izločanja modelov, ki posledično zmanjšujejo rob primerov blizu vrednosti 0, skušamo rešiti z metodo, ki ne uporablja referenčnega vektorja, kjer bi vsak primer imel svojo referenčno točko, kamor želimo približati njegov rob, temveč uporabljamo funkcijo, ki transformira vrednosti roba. Prikazujemo jo na sliki 3.2. Polna krivulja predstavlja funkcijo roba, prekinjena pa normaliziran odvod te funkcije, ki prikazuje sorazmerno utež na posamezen primer glede na njegov rob v posameznem koraku iskanja modela za izločitev. Metoda v vsakem koraku išče manjši $\sum f_m(rob)$ od obstoječega, pri čemer:

$$f_m(rob) = \begin{cases} 1 - (rob - 1)^4, & \text{če } rob \geq 0 \\ (rob + 1)^4 - 1, & \text{če } rob < 0 \end{cases} \tag{3.1}$$

Za metodo domnevamo, da ima morda dve koristni lastnosti. Prva je zmanjšana občutljivost na šum v podatkih. Domnevamo namreč, da v območju



Slika 3.1: Distribucija OOB-roba pred in po selekciji za množico breast-cancer z uporabo referenčnega vektorja $p = 0.075$.

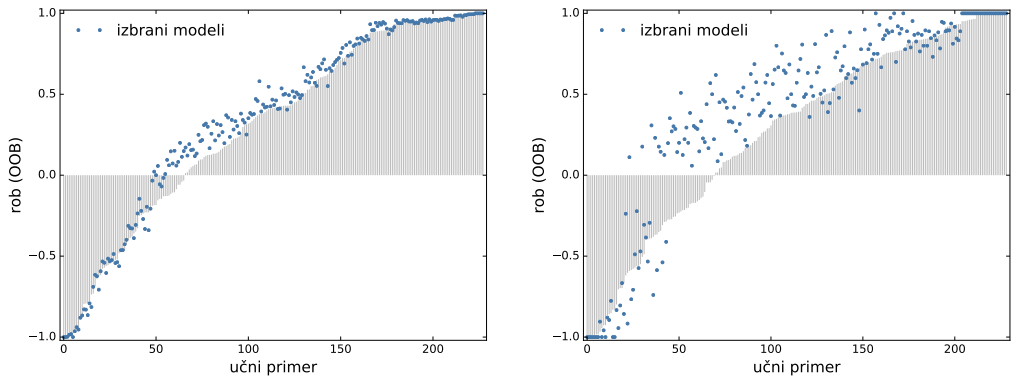


Slika 3.2: Predlagana funkcija robne razdalje, ki jo želimo maksimizirati.

z nizkim robom leži več šuma, zato se metoda osredotoči na primere z robom blizu vrednosti 0. Maksimizacija roba primerov z zelo nizkim robom je šibko utežena.

Druga koristna lastnost pa je prilagodljivost na raznolikost distribucij roba. Želeli bi namreč, da se dobro obnese tako pri problemih z visoko klasifikacijsko točnostjo kot tudi pri problemih z nizko klasifikacijsko točnostjo. Slednje dosežemo tako, da namesto minimizacije konveksne funkcije, kot pri MDSQ, maksimiziramo monotono oz. strogo naraščajočo funkcijo. S slednjim pridobimo prilagodljivost na različne distribucije robnih razdalj, saj tako vektorja robnih razdalj ne pomikamo proti vnaprej izbrani referenčni točki. Referenčna točka je namreč lahko precej daleč od povprečja robov. Pri tem pristopu vsi robovi sicer težijo k vrednosti 1, a predvidevamo, da se zaradi oblike funkcije točka ravnovesja vzpostavi že pri nižji vrednosti.

Na sliki 3.3 je prikazana distribucija OOB-roba po različnih odstotkih selekcije. Razvidno je, da metoda nima ustreznega mehanizma za zaustavljanje selekcije. V znatnem delu primerov se selekcija samodejno zaustavi šele pri nekaj odstotkih, ko pa točnost skupinskega modela večinoma že znatno upade. Zanimiva pa je ugotovitev iz slike, da se primeri z robom blizu



(a) 50 % **(b)** 5 %

Slika 3.3: Distribucija OOB-roba pri različnih odstotkih selekcije z metodo MT za podatkovno množico breast-cancer.

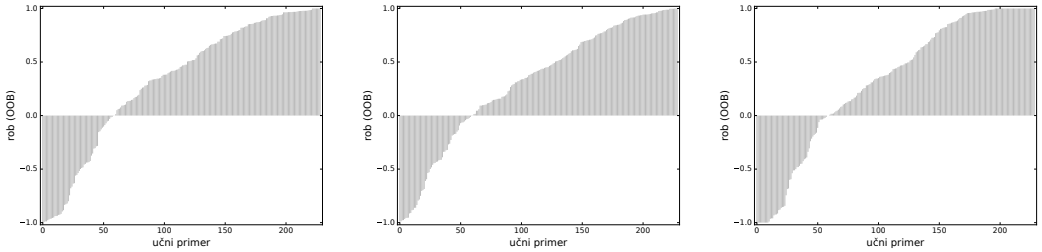
vrednosti 0 pri 5 % selekciji razdelijo na dva pola. Nekaterim se rob poveča, drugim poslabša.

Posebnost te metode, za katero težko sodimo ali je to le slabost ali kdaj tudi prednost, pa je to, da ne ohranja relativnih razmerij robov med primeri. Ker ne uporabljamo referenčnega vektorja, nima vsak rob svoje referenčne točke, h kateri bi ga približevali. Vsi robovi se v funkciji, ki jo maksimiziramo, upoštevajo le kot absolutna vrednost roba, ne pa kot razlika med robom in vrednostjo referenčnega vektorja za ta rob.

Optimalni odstotek selekcije je za posamezno vrsto skupinskega modela in podatkovno množico najbolje določiti eksperimentalno. Ugotovili smo, da preko različnih podatkovnih množic optimalna vrednost niha na celotnem testiranem območju od 1 % pa do 90 %. Zadovoljiv kompromis za privzeto vrednost je 50 % selekcija.

3.1.2 Selekcija s parametriziranim referenčnim vektorjem (PRV)

Kot alternativno metodo k MT, ki ima potencialno slabost, neohranjanje razmerij robov, smo predlagali metodo s parametriziranim vektorjem. Podobno,



(a) bagging **(b)** RF **(c)** ET

Slika 3.4: Primer distribucije OOB-roba za enak izbor učnih podatkov iz množice breast-cancer.

kot ima referenčni vektor definirana metoda MDSQ, ga lahko definiramo tudi za našo metodo, a z bistveno razliko, da nima vseh vrednostih enakih. Referenčni vektor o je določen s parametroma α in β na sledeč način:

$$o_i = \alpha + \beta * rob_i, \tag{3.2}$$

pri čemer je rob_i vrednost OOB-roba primera i pred selekcijo. Metoda v vsakem koraku algoritma požrešno išče model, ki bi zmanjšal $d(o, rob)$:

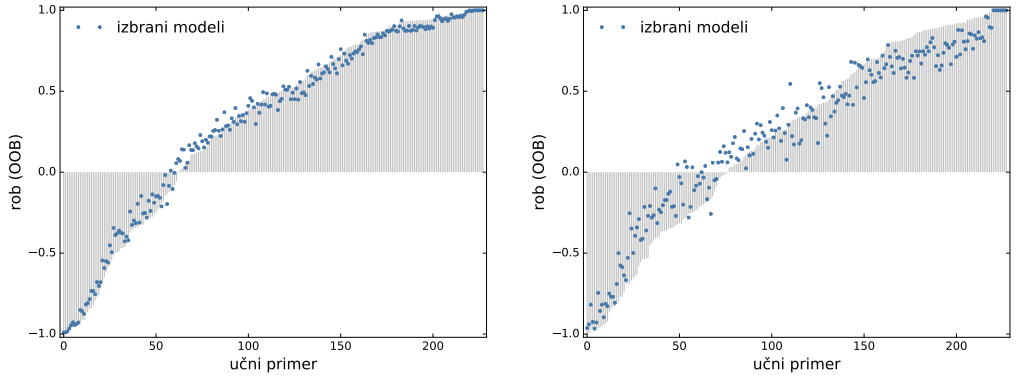
$$d(o, rob) = \sum_i (o_i - rob_i)^\gamma \tag{3.3}$$

Predlagamo območje smiselnih vrednosti parametrov $\alpha = [0, 1]$ in $\beta = [0, 1]$, $\gamma = \{2, 4\}$, pri čemer $\alpha + \beta \leq 1$.

Parametri α, β in γ so tako lahko predmet optimizacije za posamezno podatkovno množico. Distribucije robov se preko podatkovnih množic zelo razlikujejo. Izkaže se celo, da obstajajo opazne razlike med distribucijami glede na skupinski model že za enake učne množice. Kot primer na sliki 3.4 prikazujemo distribucijo roba za podatkovno množico breast-cancer.

Te ugotovitve deloma razložijo, zakaj se optimalne nastavitve parametrov referenčnega vektorja med sabo razlikujejo pri *baggingu*, RF in ET celo za iste podatkovne množice.

Točka samodejne zaustavitve selekcije, ko algoritem ne najde več modela, s katerim bi zmanjšal funkcijo d, močno niha glede na distribucijo roba in iz-



(a) 50 % **(b)** 8 % (samodejna zaustavitev)

Slika 3.5: Distribucija OOB-roba pri različnih odstotkih selekcije z metodo PRV (na osnovi *bagging* metode) za podatkovno množico breast-cancer pri $\alpha = 0.2$, $\beta = 0.5$ in $\gamma = 2$.

brano kombinacijo parametrov referenčnega vektorja. V primerjavi z metodo MT prihaja do samodejnih zaustavitev v povprečju pri višjih odstotkih selekcije, a ravno tako kot pri metodi MT ni nobenega zagotovila, da bo točka samodejne zaustavitve selekcije dosegala najboljšo točnost. Ustrezen odstotek selekcije je zato potrebno eksperimentalno določiti na učni množici ali pa poseči po vnaprej določenem relativno visokem odstotku selekcije, npr. 50 % selekciji. Na sliki 3.5 je razvidna bistveno drugačna distribucija robov po selekciji kot pri metodi MT. Razmerja med robovi se ohranjajo v večji meri, pride pa tudi do samodejne zaustavitve selekcije.

3.2 Empirično vrednotenje

3.2.1 Podatkovne množice

Smiselno je predstaviti uporabljene podatkovne množice in ključ za njihovo izbiro. Želeli smo izbrati dovolj velik nabor množic, da bi dobili kar se da reprezentativno informacijo o obnašanju predlaganih metod.

Za ponazoritve v predhodnih poglavjih in vrednotenje rezultatov smo

uporabili 34 podatkovnih množic, polovico od teh dvorazrednih, polovico večrazrednih. V izogib takšni izbiri množic, ki bi prikazovale implementirane metode v boljši luči, smo množice izbrali vnaprej in se izogibali vsakršni selektivnosti.

Izbrali smo z nekaj pojasnjenimi izjemami vse klasifikacijske množice, ki so na voljo v Orange. Izločili smo podvojeno množico (brown-selected), množici z neznano identiteto (geo-gds360, hair-eye-sex) in množice z manj kot 100 primeri zaradi velike variabilnosti rezultatov (velika odvisnosti že od naključnega deljenja množic pri prečnem preverjanju). Za lažje primerjanje rezultatov z drugimi prispevki pa smo neselektivno dodali še 4 pogosto uporabljane množice iz UCI-zbirke [27]: sonar, spambase, derm in ecoli (slednji smo dodali tudi v programski paket Orange).

V tabeli 3.1 so predstavljene podatkovne množice skupaj z viri, kolikor so avtorju poznani in navedeni v najboljši veri.

Tabela 3.1: Uporabljene podatkovne množice (levo dvorazredne, desno večrazredne)

ime	prim.	atr.[1]	atr.[2]
adult-sample[3,4]	977	14	105
bcw[3,4,5]	683	9	89
breast-cancer[3,4,6]	286	9	31
liver-disord[3,4]	345	6	6
credit-appr[3,4]	690	15	46
heart-disease[3,4,7]	303	13	25
ionosphere[3,4]	351	32	32
monks-1[3,4]	556	6	17
monks-2[3,4]	601	6	17
monks-3[3,4]	554	6	17
promoters[3,4]	106	57	228
sonar[8]	208	60	60
spambase[8]	4601	57	57
tic-tac-toe[3,4]	958	9	27
titanic[4,9]	2201	3	8
voting[3,4]	435	16	32
wdbc[3,4]	569	20	20

ime	prim.	atr.[1]	atr.[2]	razr.
anneal[3,4]	898	38	41	6
audiology[3,4,10]	226	69	154	24
balance-s[3,4]	625	4	20	3
car[3,4]	1728	6	21	4
derm[8]	366	34	35	6
ecoli[8]	336	7	9	8
hayes-roth[3,4]	132	4	15	6
horse-colic[3,4]	368	20	57	3
glass[3,4]	214	9	9	8
iris[3,4]	150	4	4	3
lymph[3,4,6]	148	18	47	4
primary[3,4,6]	339	17	37	21
shuttle[4,11]	253	6	16	8
vehicle[3,4,12]	846	18	18	4
wine[3,4]	178	13	13	3
yeast-br[4,13,14]	186	79	79	3
zoo[3,4]	101	16	36	7

[1] Atributov v originalni množici.

[2] Število atributov po preslikavi večvrednostnih v dvovrednostne (potrebno zaradi predpostavke numeričnih atributov v scikit-learn).

[3] Izvorni vir: UCI Machine Learning Repository [27].

[4] Pridobljeno iz: Orange [14], https://github.com/biolab/orange3/tree/master/Orange/datasets.

[5] O. L. Mangasarian in W. H. Wolberg. Cancer diagnosis via linear programming. 1990.

[6] Izvorno pridobljeno od: Univerzitetni klinični center, Onkološki inštitut, Ljubljana, Jugoslavija. Zahvala M. Zwitterju in M. Soklicu za pridobitev podatkov.

[7] Hungarian Institute of Cardiology. Budapest: Andras Janosi, M.D.; University Hospital, Zurich, Switzerland: William Steinbrunn, M.D.; University Hospital, Basel, Switzerland: Matthias Pfisterer, M.D.; V.A. Medical Center, Long Beach and Cleveland Clinic Foundation: Robert Detrano, M.D., Ph.D.

[8] Pridobljeno iz: UCI Machine Learning Repository [27].

[9] Dawson, Robert J. MacG. (1995), The 'Unusual Episode' Data Revisited. Journal of Statistics Education, 3., https://www.amstat.org/publications/jse/v3n3/datasets.dawson.html, osnovano na podatkih: Report on the Loss of the 'Titanic' (S.S.) (1990), British Board of Trade Inquiry Report (reprint), Gloucester, UK: Allan Sutton Publishing.

[10] Lastnik: Professor Jergen na Baylor College of Medicine.

[11] Opis: UCI Machine Learning Repository [27], natančen izvor vira iz Orange [14] neznan.

[12] Turing Institute, Glasgow, Scotland.

[13] Informacije o izboru: http://www.biolab.si/supp/bi-vizrank/yeast.htm

[14] M. B. Eisen, P. T. Spellman, P. O. Brown, in D. Botstein. Cluster analysis and display of genome-wide expression patterns. Proceedings of the National Academy of Sciences, 1998.

3.2.2 Eksperimentalni pogoji

Pogoji testiranja imajo velik vpliv na eksperimentalne rezultate selekcije v skupinskih modelih. Pri raziskovanju za prvotno začrtano zasnovo diplomskega dela smo se osredotočili na uteženo glasovanje v skupinskih modelih. Deloma je ideja temeljila na ugotovitvi Robnik Šikonje v [39], kjer predstavi uspešno metodo uteženega glasovanja, pri čemer pa pojasni, da poskusi s selekcijo dreves niso bili uspešni. Najprej smo razvili nekaj metod za uteženo glasovanje, ki so eksperimentalno pokazale izjemne rezultate. Izkazalo se je, da celo najboljše takrat, ko so bile uteži glasovanja zelo neuravnotežene, pri čemer se je vpliv nekaterih modelov praktično zanemaril in se je uteženo glasovanje izrodilo v selekcijo.

Ko smo eksperimentalni del dokončevali in poskusili za primerjavo vključiti še druge metode iz literature, pa se je izkazalo, da imajo naši rezultati osnovnega skupinskega modela v primerjavi z nekaterimi objavami v literaturi v večini primerov znatno slabšo točnost. Točnost zmanjšanih skupinskih modelov s selekcijo pa se jim je približala, a jih v le redkih primerih tudi presegla. Analiza je pokazala, da imajo naključni gozdovi v Orange [14], ki smo ga uporabljali prvotno, hrošča – privzeto nastavljene parametre za omejitev največje globine dreves in števila končnih vozlišč. Odstranitev teh dveh omejitev je pokazala uspešnost predlaganih metod, ki smo jih še nedavno smatrali za izjemno uspešne, v povsem drugi luči. S selekcijo dobljeni skupinski modeli so bili namreč v povprečju celo minimalno slabši kot enako veliki skupinski modeli, ki jih zgradimo brez selekcije (ali drugače povedano, če bi uporabili naključno selekcijo), še bolj pa je točnost odstopala od celotnih skupinskih modelov pred selekcijo.

Eno od naših prvotno predlaganih metod smo našli že predstavljeno v [26], kjer so Li in dr. uspešno uporabljali natančnost posameznega modela na OOB-množici kot utež za glasovanje na primerih podatkovnih množic z dodanim šumom. Opazimo lahko, da so točnosti na podatkovnih množicah bistveno nižje kot v drugih objavah. Ali je razlog za uspešnost v tem primeru res dodaten šum v podatkih, kot navajajo, ali pa gre, tako kot v našem pri-

meru za "neoptimalno" uporabo osnovnega algoritma, ali celo oboje, nismo natančneje ugotavljali. Smo pa naleteli še na nekatere druge objave, kjer smo bili do rezultatov podobno skeptični.

Bistvena je ugotovitev, da eksperimentalno ugotovljene uspešnosti selekcije izključno na neki specifični vrsti skupinskega modela ne moremo posplošiti. Mnogo težje je bilo namreč razviti pristop, ki izboljša že sicer izjemno natančne skupinske modele – v okviru *bagginga*, RF in ET torej take, ki so zgrajeni tako, da upoštevajo izsledke znanstvenih objav glede najustreznejše uporabe za čim višjo splošno pričakovano točnost.

V literaturi je moč zaslediti mnogo različnih pristopov k evalvaciji predlaganih metod. Bistvene razlike so:

- **Vrsta skupinskega modela**: v [32, 31, 35] za eksperimentalni del uporabijo *bagging*, v [45] pa RF. Naše empirične izkušnje kažejo, da ni vsaka metoda selekcije enako uspešna na vseh vrstah skupinskih modelov.
- **Algoritem za indukcijo dreves**: Najpogosteje naletimo na algoritma C4.5 in CART. Ali ima sam algoritem za indukcijo dreves bistveni vpliv na uspešnost selekcije, nismo raziskovali. Na podlagi empiričnih ugotovitev in pregleda literature pa sklepamo, da ga ima rezanje dreves. To se namreč odraža v nižji točnosti celotnega skupinskega modela, zato je kot prvo pri takem skupinskem modelu s selekcijo lažje izboljšati točnost, hkrati pa, kot pojasnimo v nadaljevanju, nekatere metode selekcije delujejo dobro le v skupinskih modelih s porezanimi drevesi. V [31] je pri metodi *bagging* navedeno, da v eksperimentalnem delu uporabljajo porezana CART-drevesa, v [45] pa pri RF uporabljajo neporezana CART-drevesa. Martinez in dr. v [31] pokažejo, da skupinski modeli pri metodi *bagging* s polno izgrajenimi drevesi dosegajo višjo točnost kot takšni s porezanimi drevesi, kar ni presenetljivo. Presenetljivo je to, da se selekcija z uporabo metode MDSQ bolje obnese pri modelih s porezanimi drevesi, kjer v povprečju preseže celo točnost neselektiranih modelov s polno izgrajenimi drevesi. Celo več – zaključijo,

da preizkušeni metodi selekcije pri skupinskih modelih s polno izgrajenimi CART-drevesi v povprečju ne moreta zmanjšati velikosti modelov brez izgube točnosti. V nadaljevanju bomo pokazali, da nam je z metodami, predlaganimi v tem delu, to uspelo.

- **Velikost skupinskega modela**: Zasledili smo eksperimentalne rezultate selekcije na manjših do srednje velikih skupinskih modelih, s 100 [31, 35, 45] ali 200 [32] notranjimi modeli. V [31] pokažejo, da po neki točki za metodo MDSQ nima več smisla povečevati velikosti skupinskega modela za selekcijo, saj se približamo asimptotični točnosti. Hkrati ugotovijo, da je za metodo MDSQ optimalni odstotek selekcije odvisen od velikosti skupinskega modela. Z velikostjo skupinskega modela upada in se za izbrane množice ustali okrog četrtine prvotne velikosti skupinskega modela.
- **Izbira selekcijske množice**: Eksperimentalne ugotovitve različnih objav so osnovane na različnih izbirah podatkov za selekcijsko množico. Eden prvih pristopov k selekciji [32] kot selekcijsko množico uporabi kar učno množico. V [35] razdelijo učno množico na dva dela, učno podmnožico in selekcijsko podmnožico. V [45] preizkusijo dva pristopa, selekcijo na sami učni množici in selekcijo z upoštevanjem robne razdalje na učni množici le za OOB-primere vsakega klasifikatorja. Martinez in dr. pa v [31] tudi analizirajo razlike med uporabo učne množice kot selekcijske množice in uporabo ločene selekcijske množice, pri čemer zaključijo, da uporaba ločene selekcijske množice ne zmore kompenzirati izgube natančnosti zaradi zmanjšanja učne množice za gradnjo modelov.

  Pomembno je opozoriti, da v primeru deljenja učne množice na učno in selekcijsko podmnožico dobimo pravično primerjavo z osnovnim skupinskim algoritmom le, če za osnovni algoritem uporabimo celotno učno množico, ne pa zmanjšane učne množice, ki se uporablja za pridobitev skupinskega modela pred selekcijo. Kot primera lahko navedemo [35] in [19]. Obakrat v predstavitvi rezultatov nastopajo selekcijske metode

v primerjavi z osnovno metodo *bagging*. Osnovna metoda *bagging* je naučena na 1/3 primerov, metode s selekcijo pa imajo pred testiranjem na voljo še dodatno 1/3 primerov v obliki validacijske oz. selekcijske množice. Odpira se vprašanje, kakšna bi bila točnost selekcijskih metod, če bi rezultate primerjali z *bagging* modeli, zgrajenimi na vseh učnih primerih, torej 2/3 množice. Zdi se, da takšen pristop namesto sposobnosti doseganja čim višje točnosti ovrednoti sposobnost selekcijske metode, da v procesu selekcije kombinira naključne modele, tako da uspe vključiti čim več znanja iz še nevidenih podatkov. S tem se ne bi zdelo nič narobe, če bi hkrati predstavili še primerjavo točnosti z modelom, zgrajenim na vseh učnih podatkih. V [19] je tako naiven pristop s selekcijo določenega odstotka na selekcijski množici najbolj točnih notranjih modelov prikazan kot relativno dober, čeprav se ob regularnih pogojih testiranja v večini primerov izkaže za neučinkovitega, saj ne zagotavlja raznolikosti modelov.

Na podlagi teh izsledkov smo za eksperimentalni del določili sledeče pogoje:

- Predlagane metode selekcije bomo ovrednotili na metodah *bagging*, RF in ET. Za prvi dve uporabimo privzeti algoritem na voljo v scikit-learn. ET pa uporablja za gradnjo dreves celotno učno množico, zato potrebujemo manjšo spremembo osnovne metode – za grajenje drevesa razdelimo učno množico na dve po razredih enakomerno naključno razdeljeni množici. 90 % primerov služi kot učna množica, 10 % pa kot OOB-selekcijska množica. Eksperimentalno smo ugotovili, da s tem v povprečju minimalno poslabšamo točnost skupinskega modela, a predvidevamo, da bo selekcija zmožna kompenzirati izgubo točnosti.
- Za gradnjo dreves uporabljamo privzeti, optimizirani CART-algoritem, ki je na voljo v scikit-learn. Zgradijo se polna, neporezana drevesa, kar nam zagotavlja visoko točnost polnih skupinskih modelov.
- Metode so bile zasnovane s ciljem doseganja čim višje točnosti. Zato se zdi pravilno, da sposobnost doseganja visoke točnosti primerjamo

z velikimi neselektiranimi skupinskimi modeli. Na podlagi izsledkov v [31] smo se odločili, da je ustrezno, če metode selekcije testiramo na skupinskih modelih s 1000 notranjimi modeli. Takšni modeli se v večini učnih množic že zelo približajo asimptotični točnosti za izbrano množico in vrsto skupinskega modela. Izbira se zdi smiselna tudi glede na povprečno točnost za vseh 34 podatkovnih množic v odvisnosti od števila notranjih modelov, ki jo prikazujemo na sliki 3.6.

- Glede na ugotovite v [31], zavoljo doseganja čim višje točnosti, množice ne delimo na učno in selekcijsko množico, temveč selekcijo izvajamo na OOB-primerih vsakega modela.
- Naredimo 25 ponovitev 5-kratnega prečnega preverjanja, kar smatramo za zadovoljivo število ponovitev, glede na to, da primerjamo modele z velikim številom notranjih modelov. Pri tem je, iz v naslednji točki navedenih razlogov, razdelitev množice v prečnem preverjanju narejena z enakim naključnim semenom za vse ponovitve. Generiranje modelov pa se izvede z različnimi naključnimi semeni za vseh 25 ponovitev.
- Za doseganje čim višje točnosti je ustrezno za RF in ET določiti optimalno vrednost K oz. $max\_features$. Želimo namreč izločiti možnost, da bi selekcijska metoda dosegala višjo točnost na račun kompenziranja neoptimalne vrednosti K. Optimalni K določimo s tabelaričnim iskanjem, podobno kot to stori Louppe v [28]. Pri tem uporabimo vrednosti $\{1, 2, ..., F\}$, pri čemer je F število atributov za podatkovno množico. Za RF in ET ločeno določimo optimalne vrednosti na modelih z 250 notranjimi modeli in 5-kratnim prečnim preverjanjem, ki ga 10-krat ponovimo in kot optimalni K izberemo tistega z najvišjo povprečno klasifikacijsko točnostjo. Pomembno je, da tega ne storimo za celotno podatkovno množico, temveč to ponovimo v vsakem posameznem koraku prečnega preverjanja – optimalno vrednost določimo le na učni podmnožici. Zakaj je to pomembno, pojasnjujeta Varma in Simon [43], ki ugotavljata, da s kalibracijo parametrov metod na celotnih učnih množicah dobimo optimistično pristranske modele. Ker je

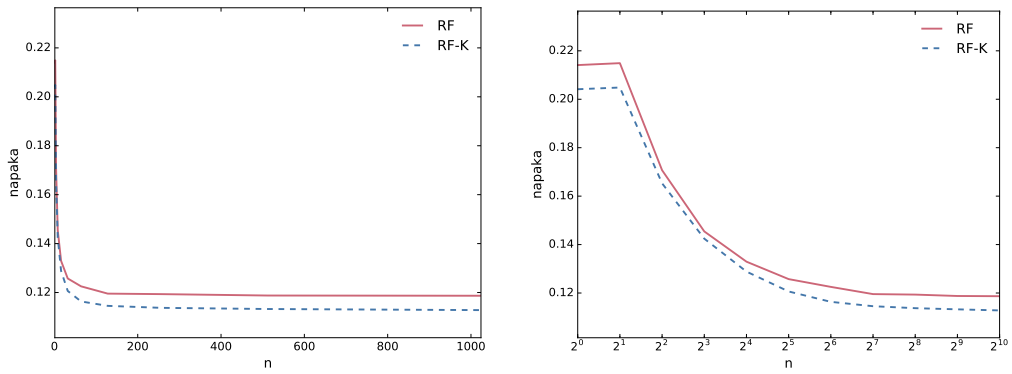
takšna kalibracija računsko zahtevna operacija, je ne moremo ponoviti za vsako od ponovitev prečnega preverjanja. Zato se, kot omenjeno v prejšnji točki, množica pri prečnem preverjanju razdeli vedno na enak način. To pomeni, da tudi pri 25 ponovitvah 5-kratnega prečnega preverjanja kalibracijo vrednosti K naredimo le 5-krat in jih lahko ponovno uporabimo.

- Brez kalibracije optimalnih vrednosti odstotka selekcije za metodi MT in PRV ter optimalnih vrednosti α, β in γ pri metodi PRV ne pričakujemo dobrih rezultatov, zato jih, podobno kot vrednost K, določimo na učni podmnožici v vsakem koraku prečnega preverjanja s tabelaričnim iskanjem. S tem se izognemo optimistično pristranskim modelom. Za spodnjo omejitev velikosti selektiranih modelov uporabimo vrednosti $\{0.1, 0.2, ..., 0.9\}$, za parametre referenčnega vektorja PRV metode pa vrednosti $\alpha = \{0.0, 0.1, ..., 1.0\}$, $\beta = \{0.0, 0.5, 0.7, 0.9\}$ in $\gamma = \{2, 4\}$, pri čemer velja omejitev $\alpha + \beta \leq 1$.
- Za nepristranskost rezultatov ne smemo ponavljati testiranj in izbirati najboljših, kljub temu da uporabljamo relativno visoko, 25-kratno ponovitev prečnega preverjanja. Temu se najlažje izognemo tako, da testiranje izvedemo samo enkrat in predstavimo dobljeni rezultat.

3.3 Rezultati

Kot priporoča Demšar [13], smo za primerjavo metod uporabili Nemenyijev test in rezultate prikazali na grafu rangov s kritičnimi razdaljami. Nemenyijev test pove, za kakšno razdaljo v rangih se morajo razlikovati metode, da jih lahko smatramo za statistično značilno različne ($p < 0.05$).

Hkrati prikazujemo tudi število zmag, remijev in porazov za pare metod, kar uporabimo za test znakov. Tudi tega podrobneje opisuje Demšar [13]. Za potrebe naših rezultatov lahko navedemo, da je metoda pri 34 podatkovnih množicah statistično značilno ($p < 0.05$) boljša od druge, če zmaga v vsaj 23 primerih, pri čemer remije enakomerno razdelimo med metodi. Uporaba



(a) linearno merilo **(b)** logaritemsko merilo

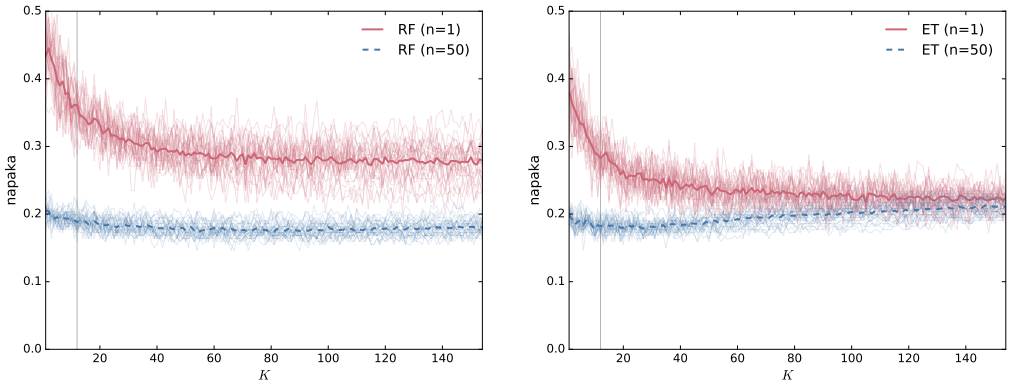
Slika 3.6: Povprečna točnost za 34 podatkovnih množic v odvisnosti od števila notranjih modelov za RF s privzetim K (RF) in optimiziranim K (RF-K).

testa znakov za navzkrižno primerjanje vseh testiranih metod ni smiselna [13], ga pa vključujemo za primerjavo parov metod s selekcijo in brez nje.

3.3.1 Vpliv parametra K

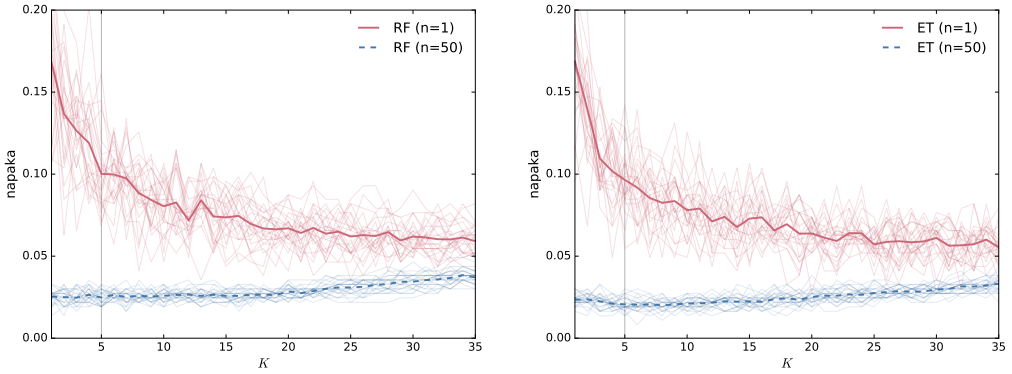
Vpliv parametra K na točnost RF in ET velikosti 1 in 50 za različne podatkovne množice lahko demonstriramo na sliki 3.7. Poudarjeni krivulji sta povprečje napak 25-krat ponovljenega 10-kratnega prečnega preverjanja, pri čemer so za vsako od 25 ponovitev množice za prečno preverjanje naključno izvlečene. Razvidno je, da imajo posamezna drevesa z manjšanjem velikosti naključnih podmnožic razpoložljivih atributov za razmejitev večjo klasifikacijsko napako, a pri tem na nekaterih problemih uspejo kot skupinski model dosegati višjo točnost. Navpična črta označuje privzeto vrednost atributa K v orodju scikit-learn, $\sqrt{|A|}$, pri čemer je A množica atributov.

Vpliv optimizacije parametra K ponazarjamo z grafom rangov na sliki 3.8 (10-krat ponovljeno 5-kratno prečno preverjanje pri 250 notranjih modelih). Testirali smo RF in ET metodi s privzeto vrednostjo K (RF, ET), RF in ET metodi z vrednostjo K, optimizirano na učni množici posameznega koraka prečnega preverjanja, (RF-K, ET-K) ter RF in ET metodi z vrednostjo K optimizirano na celotni podatkovni množici (RF-K*, ET-K*).



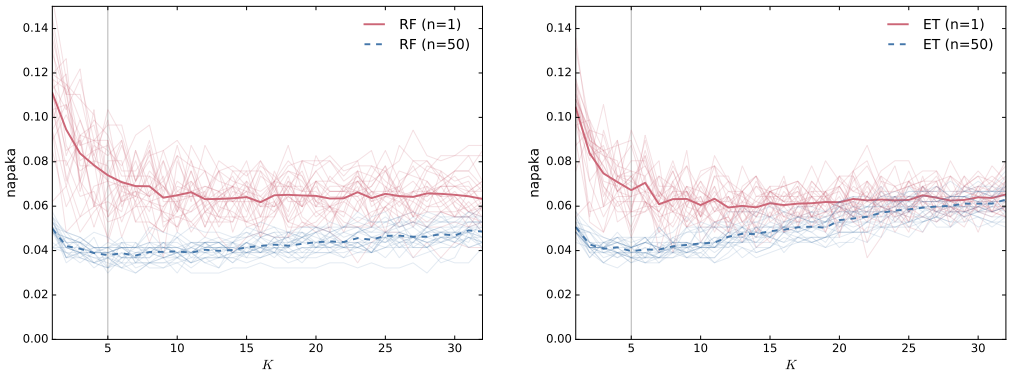
(a) audiology (RF)

(b) audiology (ET)



(c) derm (RF)

(d) derm (ET)



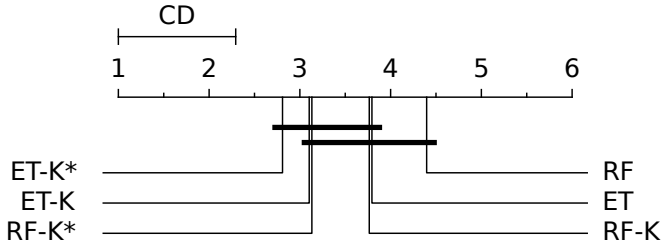
(e) voting (RF)

(f) voting (ET)

Slika 3.7: Klasifikacijska napaka RF in ET v odvisnosti od K za različne podatkovne množice.

Tabela 3.2: Vpliv kalibracije parametra K (zmage/remiji/porazi in rangi).

	RF	RF-K	RF-K*	ET	ET-K	ET-K*
RF	-					
RF-K	19/3/12	-				
RF-K*	**22/4/8**	19/5/10	-			
ET				-		
ET-K				**22/2/10**	-	
ET-K*				**23/2/9**	16/5/13	-
povpr. rang	4.397	3.132	3.765	3.794	2.809	3.103



Slika 3.8: Graf rangov (vpliv kalibracije parametra K).

Zanimal nas je odgovor na dve vprašanji. Prvo je, kakšen vpliv ima optimizacija parametra na rezultat v primerjavi s privzeto vrednostjo. Drugo pa, ali s kalibracijo parametra na celotni množici podatkov res dobimo pristransko oceno točnosti.

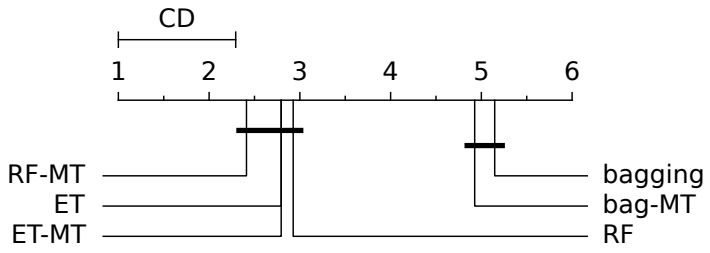
Izkaže se, da s kalibracijo vrednosti K ne dobimo statistično značilno boljše metode v primerjavi z osnovno, je pa vpliv na rezultat znaten.

Kalibriranje parametra na celotni podatkovni množici ne da izrazito boljšega modela od modela, ki ga zgradimo pravilno (s kalibracijo le na učni množici v vsakem posameznem koraku prečnega preverjanja), a je razlika kljub temu opazna. Sklepamo lahko, da bi podobno optimistično pristransko oceno točnosti dobili tudi, če bi v nadaljevanju kalibrirali parametre za selekcijo na celotni učni množici. Zato bo v nadaljevanju za določanje parametrov metod uporabljena učna množica v vsakem izmed korakov prečnega preverjanja.

Optimalne vrednosti K, izračunane s prečnim preverjanjem na celotnih

Tabela 3.3: Primerjava točnosti 25-krat ponovljenega 5-kratnega prečnega preverjanja pri selekcijski metodi MT za optimizirani K (zmage/remiji/porazi, rangi, velikost).

	bagging	bag-MT	RF	RF-MT	ET	ET-MT
bagging	-					
bag-MT	15/6/13	-				
RF			-			
RF-MT			19/4/11	-		
ET					-	
ET-MT					13/6/15	-
povpr. rang	5.147	4.926	2.926	2.412	2.794	2.794
povpr vel.	100 %	48.2 %	100 %	47.6 %	100 %	48.2 %



Slika 3.9: Primerjava *bagginga*, RF, ET (pri uporabi optimiziranih vrednosti K) brez in s selekcijsko metodo MT.

podatkovnih množicah, prikazujemo v tabeli B.1 v dodatku B.

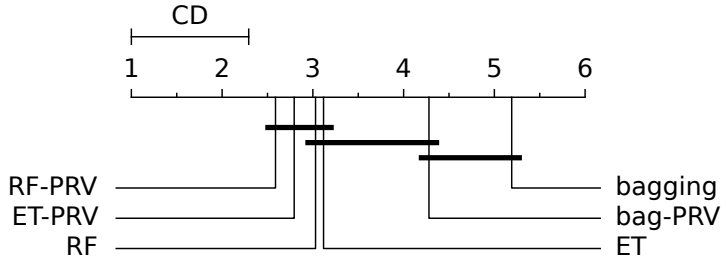
3.3.2 Selekcija

Rezultate za metodo MT (pri uporabi optimiziranih parametrov K in odstotka selekcije) navajamo v tabeli 3.3 in na pripadajočem grafu rangov na sliki 3.9. Rezultat selekcije so zmanjšani modeli, ki se v povprečju v primerjavi z večjimi obnesejo nekoliko bolje, a razlike niso statistično značilne po nobenem od testov. Podrobnejše rezultate in velikosti modelov prikazujemo v tabelah A.1 in A.2 v dodatku A.

Za metodo PRV so rezultati (pri uporabi optimiziranih parametrov K in odstotka selekcije) prikazani v tabeli 3.4 in na pripadajočem grafu rangov na

Tabela 3.4: Primerjava točnosti 25-krat ponovljenega 5-kratnega prečnega preverjanja pri selekcijski metodi PRV za optimizirani K (zmage/remiji/porazi, rangi, velikost).

	bagging	bag-PRV	RF	RF-PRV	ET	ET-PRV
bagging	-					
bag-PRV	**26/1/7**	-				
RF			-			
RF-PRV			17/2/15	-		
ET					-	
ET-PRV					18/5/11	-
povpr. rang	5.191	4.279	3.029	2.588	3.118	2.794
povpr. vel.	100 %	29.7 %	100 %	31.4 %	100 %	34.6 %



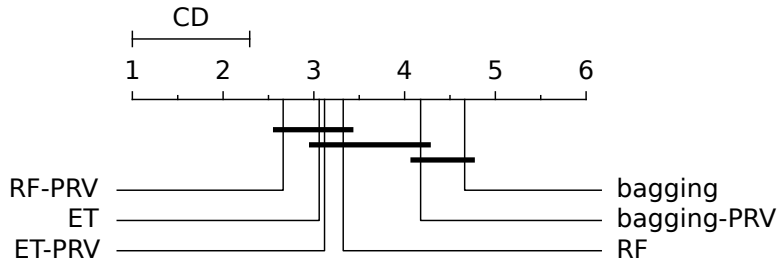
Slika 3.10: Primerjava *bagginga*, RF, ET (pri uporabi optimiziranih vrednosti K) brez in s selekcijsko metodo PRV.

sliki 3.10. Podrobnejše rezultate in velikosti modelov prikazujemo v tabelah A.3 in A.4 v dodatku A. Rezultati so v povprečju boljši kot pri metodi MT. Po testu znakov smo dobili statistično značilno boljši model s selekcijo pri metodi *bagging*. Hkrati so dobljeni modeli v povprečju manjši kot pri metodi MT.

Uspeli smo potrditi, da je metoda selekcije PRV uspešna pri optimiziranem parametru K. Ker pa se je izkazalo, da smo pri *baggingu*, kjer optimizacije tega parametra ni, dobili izjemno dober rezultat, nas je zanimalo ali bi tudi za RF in ET dobili bistveno boljši rezultat v primeru, da ne bi optimizirali parametra K. Rezultati za takšno selekcijo so prikazani v tabeli 3.5 in na pripadajočem grafu rangov na sliki 3.11.

Tabela 3.5: Primerjava točnosti 25-krat ponovljenega 5-kratnega prečnega preverjanja pri selekcijski metodi PRV za privzeti K (zmage/remiji/porazi, rangi, velikost).

	bagging	bag-PRV	RF	RF-PRV	ET	ET-PRV
bagging	-					
bag-PRV	**22/2/10**	-				
RF			-			
RF-PRV			**21/4/9**	-		
ET					-	
ET-PRV					14/3/17	-
povpr. rang	4.662	4.176	3.324	2.662	3.059	3.118
povpr. vel.	100 %	28.6 %	100 %	26.4 %	100 %	29.5 %



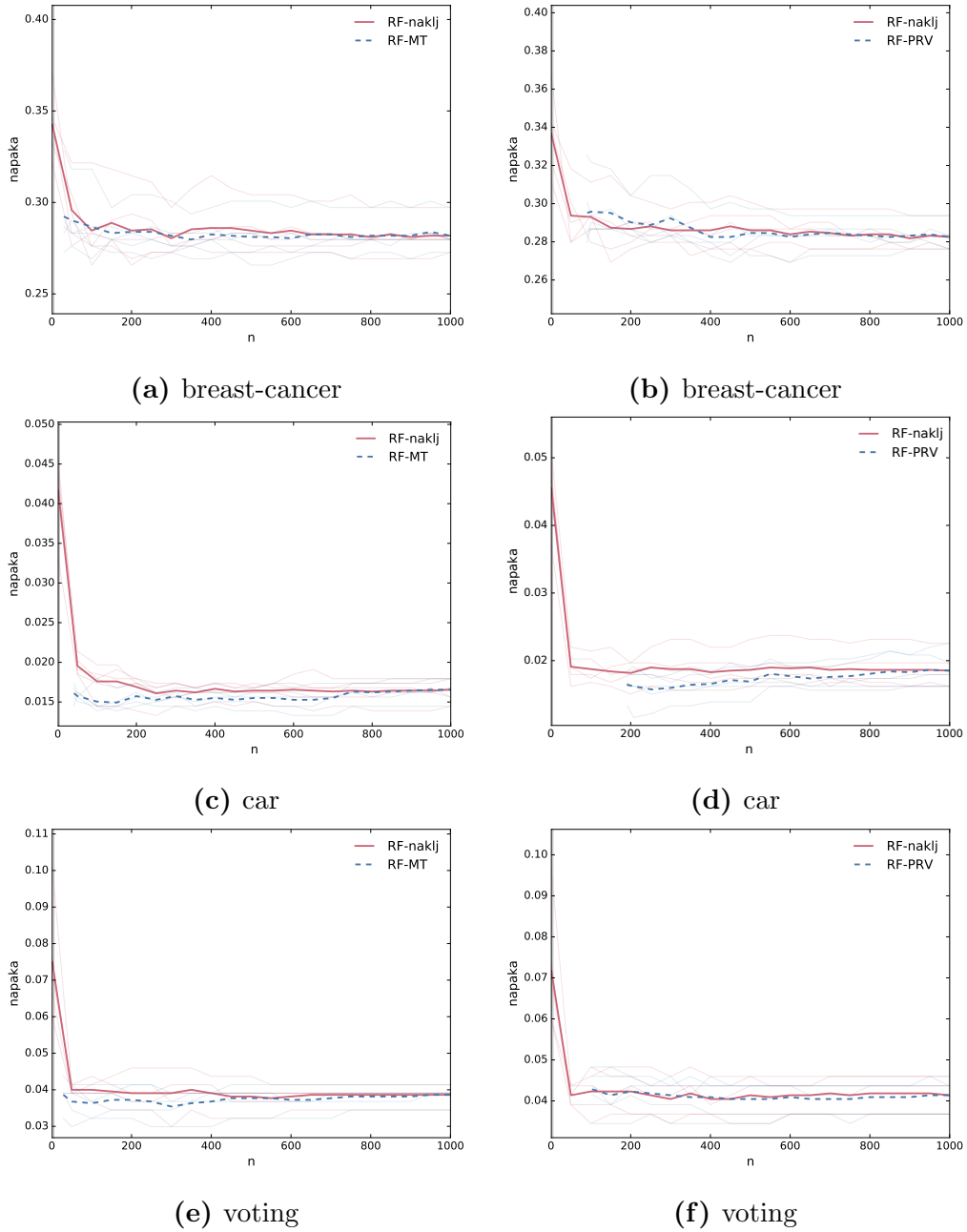
Slika 3.11: Primerjava *bagginga*, RF, ET (pri uporabi privzetih vrednosti K) brez in s selekcijsko metodo PRV.

Izkazalo se je, da smo s tem dobili statistično značilen rezultat po testu znakov tudi za selekcijo pri metodi RF, medtem ko je rezultat pri metodah *bagging* in ET nekoliko slabši, kot v rezultatih z optimizirano vrednostjo K. Razlike pripisujemo naključnosti v postopku testiranja. Kljub temu, da nismo neposredno primerjali selekcije pri neoptimizirani in optimizirani vrednosti K, se zdi varno sklepati, da selekcija ne more nadomestiti kalibracije parametra K, temveč dosega izboljšanje rezultatov na drugačen način, zato jo je ustrezno z njo kombinirati.

Na sliki 3.12 za ponazoritev delovanja metod prikazujemo napovedno napako v odvisnosti od velikosti skupinskega modela za 5 ponovitev (poudarjena krivuljo predstavlja povprečno napako). Prikazana je naključna selekcija oz. gradnja manjšega skupinskega modela in selekcija z metodama MT oz. PRV pri 1000 modelih s koraki zniževanja omejitve največje selekcije za 5 %. Za PRV uporabimo parametre $\alpha = 0.2$, $\beta = 0.7$ in $\gamma = 4$. Za takšen primer konfiguracije se selekcija pri metodi MT samodejno zaustavlja pri bistveno manjših modelih, kot pri PRV. Točnost pri majhnih modelih nekoliko upade. Do podobne ugotovitve pridemo pri večini podatkovnih množic, ki jih analiziramo na ta način. Dober rezultat je običajno možno dosegati le z določitvijo optimalnega odstotka selekcije za izbrano selekcijsko metodo in njene parametre.

Metoda MT se zdi preprostejša za uporabo, predvsem glede morebitne določitve optimalnih parametrov – določimo le odstotek selekcije. Metoda PRV pa je kompleksnejša in ponuja tudi večjo prilagodljivost za posamezno podatkovno množico, a to pomeni tudi večjo računsko zahtevnost. V praksi se metoda PRV izkaže za boljšo, saj s kalibracijo parametrov metode dosegamo tudi statistično značilno boljše rezultate. Možno bi bilo tudi pri metodi MT nastavljati obliko funkcije, kar bi morda pripeljalo do statistično značilnih razlik.

Medtem ko metodi pri nekaterih množicah dajeta dobre izboljšave napovedne točnosti, pa pri nekaterih dosegata v povprečju slabše rezultate. Ocenili bi, da razlog za to tiči v prevelikem prilagajanju učni množici. Metodi



(a) breast-cancer **(b)** breast-cancer

(c) car **(d)** car

(e) voting **(f)** voting

Slika 3.12: Primera naključne selekcije v primerjavi z metodama MT in PRV.

Tabela 3.6: Komulativni čas za izračun posamezne ponovitve prečnega preverjanja za rezultate v tabelah 3.3 in 3.4 (ločeno za osnovne metode in osnovne metode s selekcijo).

metode	čas
bagging, RF, ET (brez selekcije)	241 s
bagging, RF, ET (MT)	305 s
bagging, RF, ET (PRV)	384 s

težita k temu, da bi izbrani modeli skupaj pravilno klasificirali čim več primerov iz OOB-množice, ki v selekcijskem procesu tudi postane učna množica. Na mnogih podatkovnih množicah namreč lahko parametre izberemo tako, da uspemo večino negativnih robov OOB-primerov spremeniti v pozitivne. S tem navidez dosežemo zelo dober skupinski klasifikator, v resnici pa, podobno kot prikazujemona sliki 2.1, pridemo do prevelikega prilagajanja in s tem slabše napovedne točnosti končnega skupinskega modela.

Zaključili bi z mnenjem, da ključ do izboljšanja točnosti za nekatere množice ne leži v preprostem doseganju čim večjih vrednosti OOB-robov ali čim večjem deležu pozitivnih OOB-robov. Zdi se, da boljše rezultate dosegamo s kombinacijo povečevanja raznolikosti, pri čemer se nekoliko poveča rob primerov, ki je blizu vrednosti 0.

Če uporabimo podobno terminologijo, kot v [31], je pesimistična časovna zahtevnost obeh selekcijskih metod $O(T^2 * (1 - N_{sel}))$, kjer je T število generiranih modelov, N_{sel} pa velikost selekcijske množice. Realno je časovna zahtevnost precej nižja, saj zaradi dodatno požrešnega algoritma ne pregledujemo za vsak izločeni model celotne množice modelov, ki so še na voljo. V tabeli 3.6 navajamo čas izračunov za posamezno ponovitev prečnega preverjanja pri rezultatih, predstavljenih v tem poglavju.

Poglavje 4

Sklepne ugotovitve

Predlagali in ovrednotili smo dva učinkovita in relativno preprosta pristopa k zmanjševanju velikosti in izboljševanju natančnosti skupinskih modelov z odločitvenimi drevesi. Kot ključno ugotovitev bi izpostavili potrditev ustreznosti uporabe OOB-roba za selekcijo v skupinskih modelih. Po ugotovitvah iz literature je OOB-rob pri manjših skupinskih modelih nestabilen, saj ga pri metodi, ki uporablja vzorčenje s ponavljanjem, določa le približno 37 % dreves. Posledično doslej ni bil deležen posebne pozornosti za uporabo pri selekcijskih metodah. To težavo smo zaobšli s predlogom metod, ki pri izbiri modela za izločitev ne preiskujejo vsakič celotne množice modelov, temveč požrešno izločijo vsak model, ki ustreza kriterijski funkciji. Prednost selekcije na OOB-množici je možnost uporabe celotne razpoložljive množice podatkov za izgradnjo notranjih modelov, kar je pri manjših podatkovnih množicah še posebej pomembno za doseganje visoke točnosti. Stranski učinek metode je, da pri *baggingu* in RF izgubimo OOB-oceno točnosti skupinskega modela. Pri ET te izgube ni, ker množica že sicer ni na voljo, saj se vsi primeri uporabijo za učenje. Presenetljiva ugotovitev je, da selekcija z OOB-robom daje dobre rezultate tudi za prilagojeno metodo ET, kjer rob določa le približno 10 % dreves.

Slabost metode je neučinkovitost pri manjšem številu notranjih modelov. Zaradi uporabe OOB-roba je namreč posamezni rob določen le z manjšim

deležom vseh modelov, kar je še posebej izrazito pri metodi ET.

Z uporabo Nemenyijevega testa nismo v nobenem od primerov uspeli zaznati statistično značilnih razlik med osnovnimi metodami in metodami s selekcijo. Z uporabo testa znakov so modeli, dobljeni z metodo PRV, statistično značilno boljši pri metodi *bagging*, brez kalibracije parametra K pa tudi pri metodi RF. Pomembno je, da se kljub temu, da pri metodah selekcije dobimo v povprečju bistveno manjše modele, obnesejo primerljivo ali bolje.

Prišli smo do nepričakovane ugotovitve, da pri RF in ET optimiziranje parametra K nima bistvenega vpliva na uspešnost metod selekcije, kar smo preizkusili z metodo PRV. Pričakovali smo, da se bodo metode selekcije odrezale relativno bistveno slabše pri optimizirani vrednosti K. Sklepali smo namreč, da bo "prostor" za izboljšave manjši, a kot kaže, temu ni tako. Hkrati smo predvidevali tudi, da bomo z metodo *bagging* dosegli relativno bistveno boljšo izboljšavo kot pri RF in ET, saj bi intuitivno sklepali, da je slabšo metodo moč v večji meri izboljšati. To se izkaže za resnično pri selekcijski metodi PRV, ne pa tudi pri metodi MT.

Kot omembe vreden dosežek štejemo sposobnost metod, da izboljšajo točnost skupinskih modelov, ki uporabljajo polno zgrajena CART-drevesa. V [31] namreč empirično ugotavljajo, da so preizkušene metode v tem vidiku uspešne le, kadar uporabljajo porezana drevesa, mi pa smo pokazali, da je to možno tudi pri uporabi neporezanih dreves.

Ker smo s predlaganimi metodami in eksperimentalnim delom stremeli k preprostosti in splošnosti, obstajajo možnosti za nadaljnje raziskovanje in izboljšave. Ena izmed njih je rezanje notranjih modelov oz. dreves. To iz eksperimentalnih poskusov, ki smo jih naredili, obeta boljše rezultate, kar je skladno z ugotovitvami iz literature. Opazili smo tudi obetavne rezultate z uporabo premičnega vektorja pri metodi PRV, kjer lahko vektor iterativno pomikamo proti končni poziciji, med vsakim premikom pa izvedemo selekcijo do samodejne zaustavitve.

Razviti metodi sta dovolj splošni, da bi ju bilo možno preizkusiti tudi na drugačnih vrstah skupinskih modelov.

Dodatek A

Podrobni rezultati

V nadaljevanju so predstavljene podrobnosti nekaterih izračunov iz poglavja 3.

Tabela A.1: Podrobni rezultati k tabeli 3.3 (MT selekcija z optimiziranim K).

	bagging	bag-MT	RF	RF-MT	ET	ET-MT
adult-sample	82.52 (6)	82.56 (5)	**83.82 (1)**	83.66 (2)	83.19 (3)	83.03 (4)
bcw	95.22 (6)	95.34 (5)	97.51 (3)	97.55 (2)	97.49 (4)	**97.55 (1)**
breast-cancer	71.87 (6)	72.34 (5)	74.39 (2)	**74.43 (1)**	73.33 (4)	73.38 (3)
liver-disord	70.20 (5)	70.01 (6)	**74.06 (1)**	73.94 (2)	73.51 (4)	73.58 (3)
credit-appr	86.34 (5)	86.27 (6)	**87.97 (1)**	87.63 (2)	86.88 (3)	86.80 (4)
heart-disease	80.54 (5)	79.75 (6)	82.97 (2)	**83.05 (1)**	81.49 (4)	82.06 (3)
ionosphere	91.61 (6)	92.36 (5)	92.77 (4)	92.90 (3)	**94.75 (1)**	94.63 (2)
monks-1	**100 (3.5)**	**100 (3.5)**	**100 (3.5)**	**100 (3.5)**	**100 (3.5)**	**100.00 (3.5)**
monks-2	96.46 (5.5)	**99.15 (1.5)**	96.46 (5.5)	**99.15 (1.5)**	97.89 (4)	98.84 (3)
monks-3	97.83 (6)	**98.50 (1)**	97.83 (4)	98.14 (2)	97.83 (4)	97.83 (4)
promoters	88.15 (5)	87.32 (6)	91.17 (2)	**91.43 (1)**	90.91 (3)	90.72 (4)
sonar	80.77 (6)	80.98 (5)	82.63 (4)	83.67 (3)	**87.12 (1)**	86.73 (2)
spambase	94.41 (5)	94.40 (6)	95.52 (4)	95.52 (3)	**95.63 (1)**	95.59 (2)
tic-tac-toe	98.86 (6)	98.92 (5)	98.97 (4)	**99.11 (1)**	99.01 (3)	99.08 (2)
titanic	**79.05 (3.5)**	**79.05 (3.5)**	**79.05 (3.5)**	**79.05 (3.5)**	**79.05 (3.5)**	**79.05 (3.5)**
voting	94.34 (5)	94.29 (6)	95.47 (4)	95.77 (3)	**95.83 (1)**	95.80 (2)
wdbc	96.72 (6)	96.80 (5)	97.12 (2)	**97.37 (1)**	96.92 (4)	96.98 (3)
anneal	99.55 (6)	99.56 (5)	**99.67 (2.5)**	**99.67 (2.5)**	**99.67 (2.5)**	**99.67 (2.5)**
audiology	80.27 (6)	80.60 (5)	81.70 (3.5)	**81.98 (1)**	81.73 (2)	81.70 (3.5)
balance-s	80.17 (6)	80.31 (5)	**84.90 (1)**	84.70 (2)	83.86 (4)	84.44 (3)
car	97.81 (4.5)	97.87 (3)	97.81 (4.5)	97.88 (2)	97.71 (6)	**98.14 (1)**
derm	95.34 (5)	95.27 (6)	97.27 (3)	97.26 (4)	97.77 (2)	**97.80 (1)**
ecoli	83.68 (5)	83.42 (6)	**87.10 (1)**	86.93 (2)	86.77 (3)	86.77 (4)
hayes-roth	82.09 (5)	81.97 (6)	82.36 (3)	82.27 (4)	**83.33 (1)**	83.00 (2)
horse-colic	**71.14 (1)**	70.37 (3)	70.45 (2)	69.69 (5)	70.05 (4)	69.34 (6)
glass	76.09 (6)	76.82 (5)	**81.51 (1)**	80.88 (2)	79.36 (4)	80.09 (3)
iris	95.33 (4.5)	95.33 (4.5)	95.20 (6)	95.39 (3)	96.00 (2)	**96.16 (1)**
lymph	83.19 (5)	83.11 (6)	84.16 (2)	83.70 (4)	**84.30 (1)**	83.73 (3)
primary	42.02 (6)	42.24 (5)	44.88 (2)	**45.05 (1)**	43.83 (4)	43.98 (3)
shuttle	98.02 (5.5)	98.02 (5.5)	98.15 (4)	**98.42 (2)**	**98.42 (2)**	**98.42 (2)**
vehicle	75.13 (3)	74.94 (5)	74.66 (6)	75.06 (4)	**75.19 (1)**	75.18 (2)
wine	96.81 (5)	96.63 (6)	98.85 (2)	98.83 (3.5)	**98.88 (1)**	98.83 (3.5)
yeast-br	98.39 (5.5)	98.39 (5.5)	99.59 (2)	**99.66 (1)**	99.46 (3.5)	99.46 (3.5)
zoo	99.01 (5.5)	99.01 (5.5)	99.01 (3.5)	99.01 (3.5)	**99.72 (1)**	99.21 (2)
povpr. rang	5.147	4.926	2.926	2.412	2.794	2.794

Tabela A.2: Povprečne velikosti modelov v odstotkih za rezultate v tabeli 3.3 (MT selekcija z optimiziranim K).

	bag-MT	RF-MT	ET-MT
adult-sample	58.0	60.0	58.0
bcw	82.0	76.0	70.0
breast-cancer	38.0	54.0	46.0
liver-disord	74.0	74.0	76.0
credit-appr	56.0	52.9	47.4
heart-disease	56.0	40.0	52.0
ionosphere	22.5	36.6	29.7
monks-1	10.0	10.0	10.0
monks-2	10.0	10.0	10.0
monks-3	20.0	24.0	22.0
promoters	78.0	74.0	90.0
sonar	24.0	38.0	38.0
spambase	32.0	42.0	34.0
tic-tac-toe	22.0	24.0	24.0
titanic	78.0	66.0	58.0
voting	72.0	62.0	54.0
wdbc	26.0	28.0	24.0
anneal	36.0	24.0	28.3
audiology	42.0	58.0	74.0
balance-s	68.0	80.0	80.0
car	13.5	16.7	18.0
derm	74.0	72.0	74.0
ecoli	38.0	54.0	46.0
hayes-roth	84.0	82.0	84.0
horse-colic	46.0	46.0	36.0
glass	24.0	18.0	18.0
iris	26.3	28.2	44.3
lymph	80.5	61.3	62.6
primary	50.0	40.0	46.0
shuttle	52.0	34.0	30.0
vehicle	38.9	23.0	42.1
wine	50.0	58.0	60.0
yeast-br	74.0	66.0	76.0
zoo	84.0	78.0	78.0
povprečje	48.2	47.6	48.2

Tabela A.3: Podrobni rezultati k tabeli 3.4 (PRV-selekcija z optimiziranim K).

	bagging	bag-PRV	RF	RF-PRV	ET	ET-PRV
adult-sample	82.79 (3)	82.67 (4)	**83.46 (1)**	83.37 (2)	81.41 (6)	81.67 (5)
bcw	95.72 (6)	95.83 (5)	**97.64 (1)**	97.60 (4)	97.60 (3)	97.62 (2)
breast-cancer	69.19 (5)	68.87 (6)	**74.39 (1)**	73.64 (2)	72.38 (4)	72.56 (3)
liver-disord	70.48 (6)	70.75 (5)	**73.19 (1)**	72.50 (2)	71.44 (3)	71.32 (4)
credit-appr	87.40 (4)	87.48 (3)	**87.70 (1)**	87.69 (2)	85.77 (6)	85.90 (5)
heart-disease	83.55 (5)	83.43 (6)	83.88 (3)	83.96 (2)	83.72 (4)	**83.97 (1)**
ionosphere	91.72 (6)	92.02 (5)	93.06 (3)	93.00 (4)	**93.62 (1)**	93.60 (2)
monks-1	**100 (1.5)**	**100 (1.5)**	99.91 (5)	99.99 (3)	99.67 (6)	99.94 (4)
monks-2	98.15 (5.5)	99.30 (2)	98.15 (5.5)	**99.30 (1)**	99.00 (4)	99.05 (3)
monks-3	97.83 (4.5)	97.88 (2)	97.83 (4.5)	**97.91 (1)**	97.83 (4.5)	97.83 (4.5)
promoters	84.00 (6)	86.79 (5)	90.00 (4)	90.30 (3)	**91.32 (1)**	90.91 (2)
sonar	79.27 (6)	82.04 (5)	84.37 (4)	85.38 (3)	**88.96 (1)**	88.62 (2)
spambase	94.72 (6)	94.85 (5)	**95.93 (1)**	95.87 (2)	95.69 (4)	95.70 (3)
tic-tac-toe	99.07 (5)	99.09 (4)	98.98 (6)	**99.20 (1)**	99.17 (2.5)	99.17 (2.5)
titanic	78.89 (5.5)	78.91 (4)	78.89 (5.5)	78.92 (3)	**79.05 (1)**	79.04 (2)
voting	95.60 (5)	95.60 (6)	96.08 (4)	**96.25 (1)**	96.13 (3)	96.21 (2)
wdbc	95.25 (5)	95.68 (3)	95.19 (6)	95.64 (4)	96.25 (2)	**96.37 (1)**
anneal	99.55 (6)	99.66 (5)	99.67 (4)	99.75 (3)	99.88 (2)	**99.89 (1)**
audiology	83.01 (5)	82.64 (6)	**84.04 (1)**	83.73 (2)	83.31 (3)	83.06 (4)
balance-s	79.16 (6)	79.94 (5)	85.27 (2)	**85.52 (1)**	84.44 (4)	84.65 (3)
car	97.78 (5.5)	98.02 (2)	97.78 (5.5)	97.93 (4)	97.95 (3)	**98.33 (1)**
derm	95.55 (6)	95.79 (5)	97.17 (4)	97.41 (3)	**97.95 (1)**	97.66 (2)
ecoli	84.83 (6)	84.96 (5)	**88.20 (1)**	87.77 (2)	87.49 (3)	87.30 (4)
hayes-roth	82.00 (5)	**83.45 (1)**	81.64 (6)	83.30 (2)	82.58 (4)	82.82 (3)
horse-colic	70.98 (4)	71.46 (3)	**72.27 (1)**	71.72 (2)	69.44 (6)	69.68 (5)
glass	76.64 (6)	77.33 (5)	**80.19 (1)**	79.89 (2)	78.73 (3.5)	78.73 (3.5)
iris	94.67 (5.5)	**95.07 (1)**	94.83 (3)	94.67 (5.5)	94.83 (4)	94.88 (2)
lymph	82.57 (6)	82.59 (5)	84.89 (3)	84.70 (4)	**86.41 (1)**	86.11 (2)
primary	40.97 (6)	41.23 (5)	44.80 (2)	**44.99 (1)**	43.34 (4)	43.92 (3)
shuttle	98.42 (4)	98.21 (6)	**98.81 (1.5)**	**98.81 (1.5)**	98.42 (4)	98.42 (4)
vehicle	73.44 (6)	73.90 (5)	74.02 (4)	74.58 (3)	75.04 (2)	**75.31 (1)**
wine	96.09 (6)	96.16 (5)	98.52 (3)	97.96 (4)	**98.67 (1)**	98.54 (2)
yeast-br	97.51 (6)	97.59 (5)	**99.46 (2)**	**99.46 (2)**	**99.46 (2)**	99.31 (4)
zoo	**98.02 (2.5)**	97.98 (5)	**98.02 (2.5)**	97.86 (6)	**98.02 (2.5)**	**98.02 (2.5)**
povpr. rang	5.191	4.279	3.029	2.588	3.118	2.794

Tabela A.4: Povprečne velikosti modelov v odstotkih za rezultate v tabeli 3.4 (PRV-selekcija z optimiziranim K).

	bag-PRV	RF-PRV	ET-PRV
adult-sample	52.0	56.0	56.4
bcw	20.0	37.1	39.3
breast-cancer	72.0	56.4	54.5
liver-disord	44.2	46.3	53.3
credit-appr	41.0	43.8	35.1
heart-disease	68.0	74.0	65.3
ionosphere	24.0	34.0	35.5
monks-1	50.6	37.0	35.1
monks-2	10.0	10.0	11.9
monks-3	26.0	22.0	21.7
promoters	40.0	42.0	39.8
sonar	20.0	27.4	32.2
spambase	23.6	23.0	26.4
tic-tac-toe	10.0	11.4	24.5
titanic	12.0	12.0	10.0
voting	26.8	20.7	22.7
wdbc	28.0	26.0	16.2
anneal	23.3	20.8	37.1
audiology	39.3	43.8	50.6
balance-s	24.7	41.8	58.5
car	10.0	10.0	11.3
derm	24.0	22.0	20.6
ecoli	32.0	44.0	33.0
hayes-roth	12.0	14.1	21.5
horse-colic	27.8	25.3	27.3
glass	16.6	30.0	40.1
iris	20.0	18.2	36.9
lymph	26.0	36.4	47.8
primary	42.0	48.0	54.0
shuttle	36.3	41.1	56.3
vehicle	13.1	12.9	20.5
wine	20.1	23.0	22.0
yeast-br	38.4	35.9	34.6
zoo	35.2	22.9	26.0
povprečje	29.7	31.4	34.6

Dodatek B

Kalibrirani parametri

V poglavju 3 smo optimalno vrednost K za posamezne izračune določili posebej za vsako posamezno delitev podatkovne množice v postopku prečnega preverjanja. Enako velja za odstotke selekcije pri metodah MT in PRV ter parametre referenčnega vektorja α, β in γ pri metodi PRV.

Tako smo se izognili optimistično pristranskim ocenam točnosti uporabljenih modelov, ki so jim bile podane omenjene vrednosti. Lahko pa, po dobljenih ocenah točnosti modelov, zgradimo končne modele na celotnih množicah podatkov, ki jih imamo na voljo. Takšni modeli so primerni za dejanske napovedi novih primerov z neznanim razredom. Za te modele je ustrezno hkrati uporabiti na celotnih podatkovnih množicah izračunane optimalne vrednosti omenjenih parametrov. Na tem mestu pa jih navajamo za večjo popolnost predstavitve delovanja predlaganih metod.

Vrednosti so določene na enak način kot v posameznem koraku prečnega preverjanja v poglavju 3.

[1]Privzeta vrednost v scikit-learn za klasifikacijske probleme pri RF in ET.
[2]Optimalni K za RF.
[3]Optimalni K za ET.

Tabela B.1: Privzete in optimalne vrednosti K za podatkovne množice

ime	atr.	K_{sqrt}[1]	K_{RF}[2]	K_{ET}[3]
adult-sample	105	10	20	25
bcw	89	9	1	1
breast-cancer	31	5	1	1
liver-disord	6	2	1	4
credit-appr	46	6	3	4
heart-disease	25	5	2	1
ionosphere	32	5	8	6
monks-1	17	4	4	4
monks-2	17	4	17	9
monks-3	17	4	6	5
promoters	228	15	16	21
sonar	60	7	6	12
spambase	57	7	2	4
tic-tac-toe	27	5	7	6
titanic	8	2	3	4
voting	32	5	9	5
wdbc	20	4	7	8
anneal	41	6	6	13
audiology	154	12	54	16
balance-s	20	4	3	2
car	21	4	21	15
derm	35	5	4	6
ecoli	9	3	2	5
hayes-roth	15	3	6	6
horse-colic	57	7	18	16
glass	9	3	2	8
iris	4	2	3	3
lymph	47	6	4	2
primary	37	6	4	3
shuttle	16	4	8	7
vehicle	18	4	7	16
wine	13	3	1	4
yeast-br	79	8	2	3
zoo	36	6	21	16

Tabela B.2: Optimalne spodnje omejitve velikosti selektiranih modelov za metodi MT in PRV (1000 notranjih modelov, optimiziran parameter K).

ime	bag-MT	RF-MT	ET-MT	bag-PRV	RF-PRV	ET-PRV
adult-sample	0.8	0.6	0.6	0.6	0.6	0.1
bcw	0.6	0.6	0.7	0.1	0.6	0.4
breast-cancer	0.7	0.9	0.6	0.3	0.9	0.9
liver-disord	0.9	0.9	0.9	0.3	0.9	0.7
credit-appr	0.8	0.8	0.9	0.2	0.6	0.1
heart-disease	0.6	0.2	0.4	0.7	0.1	0.1
ionosphere	0.1	0.9	0.4	0.1	0.1	0.1
monks-1	0.1	0.1	0.1	0.5	0.1	0.1
monks-2	0.1	0.1	0.1	0.1	0.1	0.1
monks-3	0.1	0.1	0.5	0.2	0.2	0.1
promoters	0.8	0.6	0.4	0.3	0.5	0.7
sonar	0.9	0.5	0.9	0.1	0.1	0.1
spambase	0.2	0.2	0.7	0.1	0.1	0.8
tic-tac-toe	0.3	0.2	0.6	0.1	0.1	0.1
titanic	0.5	0.2	0.9	0.7	0.5	0.1
voting	0.2	0.3	0.9	0.6	0.1	0.5
wdbc	0.2	0.6	0.3	0.2	0.1	0.1
anneal	0.4	0.1	0.1	0.1	0.1	0.4
audiology	0.6	0.4	0.4	0.6	0.8	0.6
balance-s	0.6	0.9	0.9	0.1	0.3	0.5
car	0.1	0.1	0.2	0.1	0.1	0.1
derm	0.9	0.7	0.8	0.1	0.2	0.3
ecoli	0.1	0.4	0.6	0.2	0.4	0.5
hayes-roth	0.6	0.8	0.7	0.1	0.1	0.2
horse-colic	0.9	0.9	0.7	0.2	0.1	0.5
glass	0.4	0.6	0.2	0.1	0.2	0.3
iris	0.6	0.9	0.5	0.1	0.7	0.1
lymph	0.9	0.7	0.8	0.1	0.2	0.3
primary	0.8	0.8	0.5	0.7	0.9	0.3
shuttle	0.4	0.8	0.8	0.4	0.4	0.2
vehicle	0.7	0.1	0.1	0.1	0.1	0.1
wine	0.9	0.5	0.3	0.1	0.1	0.6
yeast-br	0.9	0.6	0.6	0.1	0.1	0.5
zoo	0.8	0.9	0.8	0.2	0.2	0.1

Tabela B.3: Optimalni parametri referenčnega vektorja za metodo PRV (1000 notranjih modelov, optimiziran parameter K).

ime	α_{bag}	β_{bag}	γ_{bag}	α_{RF}	β_{RF}	γ_{RF}	α_{ET}	β_{ET}	γ_{ET}
adult-sample	0.0	0.0	2	0.9	0.0	2	0.3	0.5	4
bcw	0.2	0.0	4	0.0	0.9	4	0.4	0.5	4
breast-cancer	0.2	0.5	2	0.0	0.7	4	0.0	0.0	2
liver-disord	0.0	0.0	2	0.0	0.9	4	0.1	0.9	4
credit-appr	0.0	0.0	4	0.0	0.7	4	0.3	0.7	2
heart-disease	0.1	0.0	4	0.6	0.0	2	0.1	0.9	2
ionosphere	0.9	0.0	2	1.0	0.0	2	0.6	0.0	2
monks-1	0.0	0.0	2	0.7	0.0	2	0.8	0.0	2
monks-2	0.8	0.0	4	1.0	0.0	4	0.9	0.0	2
monks-3	1.0	0.0	2	1.0	0.0	2	0.0	0.0	2
promoters	0.0	0.0	2	0.3	0.0	4	1.0	0.0	2
sonar	0.3	0.0	2	0.4	0.0	4	0.6	0.0	2
spambase	0.8	0.0	2	0.3	0.7	2	0.4	0.0	4
tic-tac-toe	0.7	0.0	2	0.7	0.0	4	0.7	0.0	4
titanic	0.8	0.0	2	0.8	0.0	4	0.7	0.0	4
voting	0.3	0.0	2	0.1	0.9	4	0.3	0.7	2
wdbc	0.4	0.0	2	0.8	0.0	2	0.5	0.5	2
anneal	0.1	0.9	4	1.0	0.0	2	0.5	0.5	2
audiology	0.2	0.7	2	0.0	0.0	4	0.1	0.0	2
balance-s	0.9	0.0	4	0.5	0.0	4	0.2	0.5	4
car	1.0	0.0	4	0.5	0.5	4	0.4	0.0	2
derm	0.0	0.5	2	0.1	0.0	4	1.0	0.0	4
ecoli	0.0	0.0	2	0.0	0.0	4	0.9	0.0	2
hayes-roth	0.7	0.0	4	0.8	0.0	4	0.1	0.9	4
horse-colic	0.1	0.0	4	0.0	0.9	2	0.9	0.0	2
glass	1.0	0.0	4	0.4	0.5	2	1.0	0.0	4
iris	0.0	0.5	4	0.0	0.0	4	1.0	0.0	2
lymph	0.2	0.0	4	0.3	0.0	2	0.2	0.0	2
primary	0.5	0.5	4	0.9	0.0	2	1.0	0.0	2
shuttle	0.0	0.0	2	0.4	0.5	4	0.0	0.0	4
vehicle	0.3	0.0	4	0.0	0.0	2	0.8	0.0	2
wine	0.3	0.0	2	0.3	0.7	4	0.3	0.7	4
yeast-br	0.2	0.0	2	0.0	0.9	2	0.7	0.0	2
zoo	0.0	0.7	2	0.0	0.0	2	0.5	0.0	2

Literatura

[1] Yali Amit, Donald Geman, in Kenneth Wilder. Joint Induction of Shape Features and Tree Classifiers. *IEEE Trans. Pattern Anal. Mach. Intell.*, 19(11):1300–1305, November 1997. Dostopno na `http://dx.doi.org/10.1109/34.632990`.

[2] Mohammad A. Bagheri, Qigang Gao, in Sergio Escalera. A Framework towards the Unification of Ensemble Classification Methods. Objavljeno v *Machine Learning and Applications (ICMLA), 2013 12th International Conference on*, del 2, strani 351–355. IEEE, December 2013. Dostopno na `http://dx.doi.org/10.1109/icmla.2013.147`.

[3] Robert E. Banfield, Lawrence O. Hall, Kevin W. Bowyer, in W. Philip Kegelmeyer. Ensemble diversity measures and their application to thinning. *Information Fusion*, 6:49–62, 2005.

[4] Vasudha Bhatnagar, Manju Bhardwaj, Shivam Sharma, in Sufyan Haroon. Accuracy–diversity based pruning of classifier ensembles. *Progress in Artificial Intelligence*, 2(2):97–111, 2014. Dostopno na `http://dx.doi.org/10.1007/s13748-014-0042-9`.

[5] L. Breiman, J. H. Friedman, R. A. Olshen, in C. J. Stone. *Classification and Regression Trees*. Chapman & Hall, New York, 1984.

[6] Leo Breiman. Bagging Predictors. *Mach. Learn.*, 24(2):123–140, Avgust 1996. Dostopno na `http://dx.doi.org/10.1023/A:1018054314350`.

[7] Leo Breiman. Out-Of-Bag Estimation, 1996. Dostopno na `https://www.stat.berkeley.edu/~breiman/OOBestimation.pdf`.

[8] Leo Breiman. Pasting Small Votes for Classification in Large Databases and On-Line. *Machine Learning*, 36(1/2):85–103, 1999.

[9] Leo Breiman. Random Forests. *Mach. Learn.*, 45(1):5–32, Oktober 2001. Dostopno na `http://dx.doi.org/10.1023/A:1010933404324`.

[10] Gavin Brown in Ludmila I. Kuncheva. *"Good" and "Bad" Diversity in Majority Vote Ensembles*, strani 124–133. Springer Berlin Heidelberg, Berlin, Heidelberg, 2010. Dostopno na `http://dx.doi.org/10.1007/978-3-642-12127-2_13`.

[11] Tom Bylander. Estimating Generalization Error on Two-Class Datasets Using Out-of-Bag Estimates. *Machine Learning*, 48(1):287–297, 2002. Dostopno na `http://dx.doi.org/10.1023/A:1013964023376`.

[12] Tom Bylander in Dennis Hanzlik. *Estimating generalization error using out-of-bag estimates*, strani 321–327. AAAI, 1 1999.

[13] Janez Demšar. Statistical Comparisons of Classifiers over Multiple Data Sets. *J. Mach. Learn. Res.*, 7:1–30, December 2006. Dostopno na `http://www.jmlr.org/papers/v7/demsar06a`.

[14] Janez Demšar, Tomaž Curk, Aleš Erjavec, Črt Gorup, Tomaž Hočevar, Mitar Milutinovič, Martin Možina, Matija Polajnar, Marko Toplak, Anže Starič, Miha Štajdohar, Lan Umek, Lan Žagar, Jure Žbontar, Marinka Žitnik, in Blaž Zupan. Orange: Data Mining Toolbox in Python. *Journal of Machine Learning Research*, 14:2349–2353, 2013. Dostopno na `http://jmlr.org/papers/v14/demsar13a.html`.

[15] Thomas G. Dietterich. Ensemble Methods in Machine Learning. Objavljeno v *Proceedings of the First International Workshop on Multiple Classifier Systems*, MCS '00, strani 1–15, London, UK, UK, 2000. Springer-Verlag.

[16] Saso Džeroski in Bernard Ženko. Is Combining Classifiers with Stacking Better than Selecting the Best One? *Machine Learning*, 54(3):255–273, 2004. Dostopno na `http://dx.doi.org/10.1023/B:MACH.0000015881.36452.6e`.

[17] Jerome H. Friedman in Werner Stuetzle. Projection pursuit regression. *Journal of the American Statistical Association*, 76:817–823, 1981.

[18] Pierre Geurts, Damien Ernst, in Louis Wehenkel. Extremely Randomized Trees. *Mach. Learn.*, 63(1):3–42, apr 2006. Dostopno na `http://dx.doi.org/10.1007/s10994-006-6226-1`.

[19] Li Guo in Samia Boukir. Margin-based Ordered Aggregation for Ensemble Pruning. *Pattern Recogn. Lett.*, 34(6):603–609, April 2013. Dostopno na `http://dx.doi.org/10.1016/j.patrec.2013.01.003`.

[20] Trevor J. Hastie, Robert John Tibshirani, in Jerome H. Friedman. *The elements of statistical learning : data mining, inference, and prediction.* Springer series in statistics. Springer, New York, 2009. Autres impressions : 2011 (corr.), 2013 (7e corr.). Dostopno na `http://opac.inria.fr/record=b1127878`.

[21] Daniel Hernández-Lobato, Gonzalo Martínez-Muñoz, in Alberto Suárez. Empirical analysis and evaluation of approximate techniques for pruning regression bagging ensembles. *Neurocomputing*, 74(12–13):2250 – 2264, 2011. Dostopno na `http://www.sciencedirect.com/science/article/pii/S0925231211001068`.

[22] Tin Kam Ho. The Random Subspace Method for Constructing Decision Forests. *IEEE Trans. Pattern Anal. Mach. Intell.*, 20(8):832–844, Avgust 1998. Dostopno na `http://dx.doi.org/10.1109/34.709601`.

[23] ALBERT HUNG-REN KO, ROBERT SABOURIN, in ALCEU DE SOUZA BRITTO. COMPOUND DIVERSITY FUNCTIONS FOR ENSEMBLE SELECTION. *International Journal of Pattern Recognition*

and Artificial Intelligence, 23(04):659–686, 2009. Dostopno na `http://www.worldscientific.com/doi/abs/10.1142/S021800140900734X`.

[24] Igor Kononenko in Matjaz Kukar. *Machine Learning and Data Mining: Introduction to Principles and Algorithms*. Horwood Publishing Limited, 2007.

[25] Ludmila I. Kuncheva in Christopher J. Whitaker. Measures of Diversity in Classifier Ensembles and Their Relationship with the Ensemble Accuracy. *Machine Learning*, 51(2):181–207, 2003. Dostopno na `http://dblp.uni-trier.de/db/journals/ml/ml51.html#KunchevaW03`.

[26] Hong Bo Li, Wei Wang, Hongwei Ding, in Jin Dong. Trees Weighting Random Forest Method for Classifying High-Dimensional Noisy Data. Objavljeno v *IEEE 7th International Conference on e-Business Engineering, ICEBE 2010, Shanghai, China, November 10-12, 2010*, strani 160–163, 2010. Dostopno na `http://dx.doi.org/10.1109/ICEBE.2010.99`.

[27] M. Lichman. UCI Machine Learning Repository, 2013. Dostopno na `http://archive.ics.uci.edu/ml`.

[28] Gilles Louppe. *Understanding Random Forests: From Theory to Practice*. Doktorska dizertacija, University of Liege, Belgium, 10 2014. arXiv:1407.7502. Dostopno na `http://arxiv.org/abs/1407.7502`.

[29] Dragos D. Margineantu in Thomas G. Dietterich. Pruning Adaptive Boosting. Objavljeno v *Proceedings of the Fourteenth International Conference on Machine Learning*, ICML '97, strani 211–218, San Francisco, CA, USA, 1997. Morgan Kaufmann Publishers Inc.

[30] Gonzalo Martínez-Muñoz in Alberto Suárez. Pruning in Ordered Bagging Ensembles. Objavljeno v *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, strani 609–616, New York, NY, USA, 2006. ACM.

[31] Gonzalo Martínez-Muñoz, Daniel Hernández-Lobato, in Alberto Suárez. An Analysis of Ensemble Pruning Techniques Based on Ordered Aggregation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 31(2):245–259, 2009. Dostopno na `http://dblp.uni-trier.de/db/journals/pami/pami31.html#Martinez-MunozHS09`.

[32] Gonzalo Martínez-Muñoz in Alberto Suárez. Aggregation ordering in bagging. Objavljeno v *Proc. of the IASTED International Conference on Artificial Intelligence and Applications*, strani 258–263. Acta Press, 2004.

[33] J. N. Morgan in J. A. Sonquist. Problems in the Analysis of Survey Data, and a Proposal. *Journal of the American Statistical Association*, 58:415–435, 1963.

[34] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, in E. Duchesnay. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[35] Chao Qian, Yang Yu, in Zhi-Hua Zhou. Pareto Ensemble Pruning. Objavljeno v *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, January 25-30, 2015, Austin, Texas, USA.*, strani 2935–2941, 2015. Dostopno na `http://www.aaai.org/ocs/index.php/AAAI/AAAI15/paper/view/9507`.

[36] J. R. Quinlan. Induction of Decision Trees. *Mach. Learn.*, 1(1):81–106, Marec 1986. Dostopno na `http://dx.doi.org/10.1023/A:1022643204877`.

[37] J. R. Quinlan. Bagging, Boosting, and C4.5. Objavljeno v *Proceedings of the Thirteenth National Conference on Artificial Intelligence - Volume 1*, AAAI'96, strani 725–730. AAAI Press, 1996.

[38] J. Ross Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1993.

[39] Marko Robnik-Šikonja. *Machine Learning: ECML 2004: 15th European Conference on Machine Learning, Pisa, Italy, September 20-24, 2004. Proceedings*, chapter Improving Random Forests, strani 359–370. Springer Berlin Heidelberg, Berlin, Heidelberg, 2004. Dostopno na http://dx.doi.org/10.1007/978-3-540-30115-8_34.

[40] Lior Rokach. Ensemble-based classifiers. *Artificial Intelligence Review*, 33(1):1–39, 2009. Dostopno na http://dx.doi.org/10.1007/s10462-009-9124-7.

[41] Robert E. Schapire, Yoav Freund, Peter Bartlett, in Wee Sun Lee. Boosting the margin: a new explanation for the effectiveness of voting methods. *Ann. Statist.*, 26(5):1651–1686, 10 1998. Dostopno na http://dx.doi.org/10.1214/aos/1024691352.

[42] E. K. Tang, P. N. Suganthan, in X. Yao. An Analysis of Diversity Measures. *Mach. Learn.*, 65(1):247–271, Oktober 2006. Dostopno na http://dx.doi.org/10.1007/s10994-006-9449-2.

[43] Sudhir Varma in Richard Simon. Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics*, 7(1):1–8, 2006. Dostopno na http://dx.doi.org/10.1186/1471-2105-7-91.

[44] Zongxia Xie, Yong Xu, Qinghua Hu, in Pengfei Zhu. Margin Distribution Based Bagging Pruning. *Neurocomput.*, 85:11–19, Maj 2012. Dostopno na http://dx.doi.org/10.1016/j.neucom.2011.12.030.

[45] Fan Yang, Wei hang Lu, Lin kai Luo, in Tao Li. Margin optimization based pruning for random forest. *Neurocomputing*, 94:54 – 63, 2012. Dostopno na http://www.sciencedirect.com/science/article/pii/S0925231212003396.

[46] Zhi-Hua Zhou. *Ensemble Methods: Foundations and Algorithms*. Chapman & Hall/CRC, 1st edition, 2012.

[47] Zhi-Hua Zhou, Jianxin Wu, in Wei Tang. Ensembling neural networks: Many could be better than all. *Artificial Intelligence*, 137(1–2):239 – 263, 2002. Dostopno na `http://www.sciencedirect.com/science/article/pii/S000437020200190X`.